DOCTORAL THESIS

Probabilistic modelling of linear and directional wind and wave data with applications to the marine environment

Author:
Flora E. Karathanasi

Supervisor:
Prof. Kostas A. Belibassakis

A thesis submitted to the School of Naval Architecture and Marine Engineering
for the degree of

Doctor of Engineering

Dedicated to my grandmother, Flora

"He who knows the All,
but lacks himself,
lacks everything."

Gospel of Thomas (translated by David R. Cartlidge, 1980)

Abstract

Analysing and modelling metocean variables is crucial in various scientific fields with both socio-economic and environmental impacts, such as offshore engineering and coastal morphology studies. The rapidly emerging sector of marine renewable energy is also largely based on the sound analysis and modelling of the metocean climate in a candidate area at various temporal scales (e.g. seasonal, interannual). The involved physical processes that govern the metocean environment are highly interrelated; linear metocean variables are related with either other linear or directional variables and vice versa, rendering their adequate joint description a demanding task. On the other hand, their variability aspects are of great interest for long-term planning purposes. Furthermore, such phenomena can be realized at entirely different time scales; for instance, fatigue of structures placed in the marine environment or equilibrium of a coastal zone due to erosion/accretion patterns are affected not only by the severe (extreme) environmental conditions acting for a limited time period (e.g. storm events) but also by their repeated and continuous action corresponding to a longer-term "fatigue".

The focus of this dissertation is on the analytic probabilistic modelling and assessment of linear and directional metocean random variables aiming at an integrated and feasible approach for climate modelling. This task extends to a wide spectrum of less known probabilistic approaches, such as linear-directional joint probability models, circular regression and calibration, estimation of extreme events taking into account the directional covariate, etc. The modelling approaches refer to the long-term scale, but the methods apply equally well on any time scale. Moreover, the multitude of alternative models renders the statistical decision procedure a very delicate task, since the results of this step play a decisive role in ocean and coastal engineering applications.

In the first part, the main theoretical background is presented starting with conventional linear and directional univariate and bivariate models, along with some distributions that have been recently proved to be efficient for modelling metocean characteristics, while mixtures of different univariate models are also considered. The examined bivariate models that are based on the corresponding marginal distributions and an appropriate dependence structure, are described in detail. Both dimensions (one and two) are addressed by parametric and non-parametric models. Also, the current availability of multiple data sources leads to the necessity of validating and correcting (calibrating) metocean data with emphasis on regression models that take into consideration errors in both variables and the presence of outliers, while calibration techniques are described for linear and directional features, with the latter being rarely adopted in relevant applications despite its significance. As safety, performance and economic viability of marine structures are affected by directional features, directionality is incorporated in a standard extreme value distribution in order to examine the dependence of extreme values of linear metocean variables with a directional covariate. An alternative penalised likelihood criterion is proposed to estimate the unknown parameters, which seems to be numerically stable for optimization.

In the second part of this thesis, the above statistical methods are implemented on real data sets stemming from the disciplines of climate modelling and marine renewable energy; in some case studies, new statistical measures and methodologies are proposed. Specifically, long-term time series of offshore wind speed and direction are assessed across the Mediterranean Sea in order to identify systematic patterns and reveal the general features of the wind climatology patterns, with the proposed variability measures revealing further directional attributes of the wind flow. As marine energy applications require data of high quality, linear and directional wind and wave data obtained from less reliable, but easily-accessible, data sources (satellite data, numerical results) are calibrated using in situ measurements as a reference source. To this end,

specific robust estimators for linear variables seem to systematically provide better results than the ordinary least squares for the examined locations while, after an analytic evaluation, circular calibration based on the orthogonal distance outperforms and thus is suggested to be adopted in energy assessment studies. Wind speed and direction data are thoroughly examined by means of various (parametric) univariate and bivariate models. The evaluation of the bivariate models indicates that there is inconsistency of univariate models to the bivariate case in terms of performance. Wave energy flux and mean wave direction are also modelled using both parametric and non-parametric bivariate distributions to evaluate the validity of the latter and investigate the most appropriate for ocean energy applications and optimization of the performance of wave energy devices. The extreme events of significant wave height are modelled with a directional model in order to identify the dependence of the design values with wave direction; this model should be applied for the estimation of extreme wave (and wind) loads for any non-symmetric structure operating in the marine environment.

Moving to the coastal environment, the last part addresses wave action in sediment transport modelling based on two different approaches under the perspective of frequency and amplitude of waves; in the first one, the action of individual high waves, for a short time window, is examined on a sandy beach to give insight into the impact of hydrodynamics and circulation on sediment transport processes during and after such intense sea states. In the second case, the accumulative action of waves throughout a typical year is considered studying the corresponding impacts on a sandy beach vulnerable to erosion phenomena, with a profound touristic character. For the latter approach, a cost-effective method is introduced combined with the philosophy of wave input reduction techniques; the full wave time series is substituted by representative wave conditions that are able to initiate or not grain motion. Both case studies are based on the use of a widely recognized process-based model that integrates several distinct models.

Πιθανοθεωρητική μοντελοποίηση γραμμικών και κατευθυντικών μεταβλητών ανεμολογικών και κυματικών χαρακτηριστικών με εφαρμογές στο θαλάσσιο περιβάλλον

Φλώρα Ε. Καραθανάση

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ναυπηγών Μηχανολόγων Μηχανικών
Τομέας Ναυτικής και Θαλάσσιας Υδροδυναμικής

Περίληψη

Οι μετεωρολογικές και ωκεανογραφικές μεταβλητές (*metocean variables*) διαδραματίζουν σημαντικό ρόλο σε μια σειρά από αλληλένδετες φυσικές διεργασίες που απαντώνται στο θαλάσσιο περιβάλλον. Η ακριβής γνώση των σημαντικότερων πτυχών του ανεμολογικού και κυματικού κλίματος καθώς και η εκτίμηση ακραίων γεγονότων είναι θεμελιώδους σημασίας και συμβάλλουν στο μετριασμό κινδύνων είτε αυτές αφορούν κατασκευές είτε την ασφάλεια της ανθρώπινης ζωής στη θάλασσα. Ενδεικτικά, ορισμένοι από τους τομείς εφαρμογών που είναι στενά συνδεδεμένοι με την καλή γνώση της κλιματολογίας είναι τα έργα θαλάσσιας μηχανικής (π.χ., σχεδιασμός και κατασκευή παράκτιων υποδομών), οι υπεράκτιες δραστηριότητες (π.χ., πλατφόρμες εξόρυξης πετρελαίου), η διασπορά ρύπων σε αέρα και νερό, ο προγραμματισμός πορειών πλεύσης πλοίων, το φαινόμενο διάβρωσης-απόθεσης και οι θαλάσσιες ανανεώσιμες πηγές ενέργειας (ΘΑΠΕ), που παρουσιάζουν αυξανόμενο ενδιαφέρον για ανάπτυξη τις τελευταίες δύο δεκαετίες.

Γενικά, το μετεωρολογικό και ωκεανογραφικό κλιματικό σύστημα δεν μπορεί να περιγραφεί λεπτομερώς λόγω της μη επαρκούς ή ελλιπούς γνώσης των φυσικών νόμων και των αναρίθμητων παραγόντων που επηρεάζουν τις αντίστοιχες συνιστώσες του και προκαλούν αστάθειες και μη-γραμμικότητες. Επομένως, η ανάγκη εισαγωγής πιθανοθεωρητικών εννοιών και στατιστικών μεθόδων είναι αναγκαία για την εν λόγω περιγραφή σε μια συγκεκριμένη περιοχή και χρονική περίοδο. Οι συνήθεις γραμμικές τυχαίες μεταβλητές που χρησιμοποιούνται για τον χαρακτηρισμό των ανεμολογικών και κυματικών συνθηκών είναι η ταχύτητα του ανέμου, και το σημαντικό ύψος κύματος και η μέση περίοδος κύματος, αντίστοιχα. Παρόλα αυτά, η περιγραφή αυτή σήμερα θεωρείται αρκετά ελλιπής και για το λόγο αυτό, οι αντίστοιχες κατευθυντικές μεταβλητές (δηλ. η διεύθυνση ανέμου και κύματος) θα πρέπει επίσης να συμπεριλαμβάνονται για την ολοκληρωμένη περιγραφή του ανεμολογικού και κυματικού κλίματος. Η σημασία της κατευθυντικότητας έχει επισημανθεί από πρόσφατες μελέτες τόσο για τις ΘΑΠΕ όσο και την παράκτια διάβρωση, που καθιστούν και τους δύο κύριους άξονες εφαρμογών της παρούσας διατριβής.

Βασικός στόχος της παρούσας διατριβής είναι η ανάπτυξη μιας ολοκληρωμένης, κατά το δυνατόν, προσέγγισης για την πιθανοθεωρητική μοντελοποίηση γραμμικών και κατευθυντικών μεταβλητών ανεμολογικών και κυματικών παραμέτρων. Παρόλο που η ανάπτυξη αυτή αφορά τις συγκεκριμένες παραμέτρους εντούτοις μπορεί εύκολα να καλύψει και οποιοδήποτε άλλο κατευθυντικό περιβαλλοντικό χαρακτηριστικό (π.χ., θαλάσσια ρεύματα). Για την επίτευξη αυτού του στόχου, η εργασία εκτείνεται σε ένα ευρύ φάσμα λιγότερο γνωστών πιθανοθεωρητικών προσεγγίσεων, όπως είναι τα διδιάστατα μοντέλα (παραμετρικά και μη παραμετρικά) γραμμικών και κατευθυντικών μεταβλητών, η παλινδρόμηση και διόρθωση κατευθυντικών μεταβλητών, η εκτίμηση ακραίων γεγονότων λαμβάνοντας υπόψη την κατευθυντικότητα ως συμμεταβλητή, κ.ά. Τα μαθηματικά εργαλεία και οι μεθοδολογίες που παρουσιάζονται, εστιάζονται σε συγκεκριμένες πτυχές του κλίματος που είναι είτε άγνωστες

είτε εφαρμόζονται σπάνια. Επίσης, γενικά, ενώ αναφέρονται στη μακροπρόθεσμη κλίμακα, βρίσκουν εντούτοις εξίσου καλή εφαρμογή σε οποιαδήποτε χρονική κλίμακα. Επιπρόσθετα, η ανάλυση μπορεί εύκολα να επεκταθεί και σε άλλες συναφείς περιβαλλοντικές μεταβλητές, όπως είναι το παλιρροιακό εύρος (μεταξύ πλήμμης και ρηχίας), η θερμοκρασία και η πυκνότητα αέρα και νερού, η αλατότητα και η ηλιακή ακτινοβολία, καθώς και σε άλλα πεδία των γεωεπιστημών, όπως η μετεωρολογία, η γεωλογία, η γεωγραφία και η οικολογία.

Σχετικά με την ανάλυση και μοντελοποίηση μετεωρολογικών και ωκεανογραφικών μεταβλητών, η εργασία επικεντρώνεται στους ακόλουθους στόχους:

- ανάπτυξη μιας ολοκληρωμένης προσέγγισης για την μοντελοποίηση γραμμικών και κατευθυντικών τυχαίων μεταβλητών μετεωρολογικών και ωκεανογραφικών παραμέτρων
- επισήμανση νέων χαρακτηριστικών σχετικών με την κατευθυντική μεταβλητότητα του ανεμολογικού και κυματικού κλίματος
- συστηματική μελέτη και σύγκριση διάφορων παραμετρικών μοντέλων για γραμμικές και κατευθυντικές μεταβλητές τόσο στη μονοδιάστατη όσο και τη διδιάστατη περίπτωση μέσω στατιστικών δεικτών
- διεξοδική αξιολόγηση παραμετρικών και μη παραμετρικών μοντέλων κατάλληλων για την από κοινού περιγραφή γραμμικών και κατευθυντικών μεταβλητών ανεμολογικών και κυματικών χαρακτηριστικών
- ανάπτυξη μεθοδολογίας για την αξιολόγηση λιγότερο αξιόπιστων πηγών δεδομένων που λαμβάνουν υπόψη τις έκτροπες παρατηρήσεις σε ένα δείγμα γραμμικών μεταβλητών, καθώς και η διόρθωση κατευθυντικών μεταβλητών (η οποία συνήθως δεν εφαρμόζεται στη συνηθισμένη πρακτική)
- χρήση τεχνικών από τη μονομεταβλητή ανάλυση ακραίων τιμών μέσω μοντέλων που περιλαμβάνουν την κατευθυντικότητα ως συμμεταβλητή, η εισαγωγή ενός νέου κριτηρίου ποινικοποιημένης πιθανοφάνειας για την εκτίμηση των παραμέτρων και η διερεύνηση της συμπεριφοράς του μοντέλου με βάση διαφορετικές μεθόδους επιλογής κατωφλίου και απο-ομαδοποίησης (declustering) ακραίων δεδομένων.

Η εκτίμηση και η πρόβλεψη των πεδίων της μεταφοράς ιζημάτων και των μεταβολών του θαλάσσιου πυθμένα λόγω της δράσης των κυμάτων αφορά ένα ακόμα ευρύ πεδίο εφαρμογών με ιδιαίτερο ενδιαφέρον και σε άμεση σχέση με τα ανωτέρω μαθηματικά εργαλεία που εξετάζονται. Οι μεταβαλλόμενες χρονικές κλίμακες μελέτης των κυματισμών, από την εκδήλωση καταιγιδικών συμβάντων μέσα σε κάποιες ώρες έως το τυπικό κυματικό κλίμα σε μια παράκτια περιοχή, έχουν ως αποτέλεσμα τη διαφορετική απόκριση των ιζημάτων στην παράκτια ζώνη. Για το λόγο αυτό, η μετακίνηση των ιζημάτων εξετάζεται i) λαμβάνοντας υπόψη μεμονωμένα καταιγιδικά φαινόμενα που εξελίσσονται σε ένα σύντομο χρονικό διάστημα, και ii) την επαναλαμβανόμενη δράση των κυματισμών μέσα σε ένα έτος.

Στο πλαίσιο αυτό, τέθηκαν ορισμένοι πρόσθετοι στόχοι για την παρούσα διατριβή:

- αξιολόγηση των επιπτώσεων των κυματισμών στις διεργασίες ιζηματομεταφοράς κατά τη διάρκεια και έπειτα από έντονες καταστάσεις θάλασσας
- ανάπτυξη μεθοδολογίας για την εκτίμηση και πρόβλεψη του επιπέδου του πυθμένα, η οποία είναι αποδοτικότερη ως προς τον υπολογιστικό χρόνο και βασίζεται στη φιλοσοφία των τεχνικών μείωσης κυματικών δεδομένων εισαγωγής (wave input reduction techniques), λαμβάνοντας υπόψη την αθροιστική δράση των κυμάτων.

Οι δύο ανωτέρω στόχοι αποσκοπούν στην καλύτερη κατανόηση της δυναμικής συμπεριφοράς ενός παράκτιου συστήματος, την αναγνώριση μοτίβων διάβρωσης/απόθεσης και την γρήγορη και αποδοτική πρόβλεψη αντίστοιχων μελλοντικών τάσεων. Όλες αυτές οι απόψεις είναι μεγάλης σπουδαιότητας κατά το σχεδιασμό και τη διαχείριση παράκτιων δραστηριοτήτων,

εφόσον βέβαια υπάρχει αξιολόγηση της ικανότητας του μοντέλου αναφορικά με τις τάσεις που παρατηρούνται σε σύγκριση με την πραγματική κατάσταση.

Η παρούσα εργασία χωρίζεται σε ένα εισαγωγικό κεφάλαιο, σε τρία κύρια μέρη που αποτελούν το βασικό κορμό της διατριβής, καθώς και σε ένα συμπερασματικό κεφάλαιο. Στο **εισαγωγικό μέρος** παρουσιάζεται το γενικότερο πλαίσιο της εργασίας προβάλλοντας τη σημασία των μετεωρολογικών και ωκεανογραφικών μεταβλητών σε διάφορους επιστημονικούς κλάδους και τη σημασία της μελέτης της κατευθυντικότητας στους τομείς των ΘΑΠΕ και της παράκτιας διάβρωσης. Στο πλαίσιο αυτό, περιγράφονται αναλυτικά τα κίνητρα και οι ερευνητικοί στόχοι της παρούσας εργασίας, και παρουσιάζεται η συνεισφορά καθώς και οι δημοσιεύσεις σε επιστημονικά περιοδικά και διεθνή συνέδρια, που προέκυψαν κατά τη διάρκεια αυτής της μελέτης.

Στο πρώτο μέρος, το οποίο αποτελείται από τρία κεφάλαια, αναπτύσσεται το θεωρητικό πλαίσιο για τη μοντελοποίηση γραμμικών και κατευθυντικών μεταβλητών. Πιο συγκεκριμένα, το **Κεφάλαιο 1** περιλαμβάνει τη πιθανοθεωρητική μοντελοποίηση ανεμολογικών και κυματικών μεταβλητών για το χαρακτηρισμό του κλίματος με χρήση παραμετρικών στατιστικών μοντέλων, μέσω θεωρητικών κατανομών, και μη παραμετρικών μοντέλων, που δεν απαιτούν συγκεκριμένες υποθέσεις για τη μορφή του πληθυσμού από τον οποίο προέρχεται το δείγμα. Στα πλαίσια της διατριβής, εξετάζονται αναλυτικά πολυάριθμες παραμετρικές και μη-παραμετρικές κατανομές για την από κοινού περιγραφή γραμμικών και κατευθυντικών μεταβλητών που χαρακτηρίζονται από κάποιο βαθμό εξάρτησης. Η επιλογή των κατάλληλων μονομεταβλητών και διμεταβλητών μοντέλων, παραμετρικών ή μη παραμετρικών, βασίζεται σε στατιστικά κριτήρια καλής προσαρμογής (*goodness-of-fit criteria*). Τα αποτελέσματα μιας τέτοιας ανάλυσης κρίνονται απαραίτητα στην περίπτωση μελέτης του κλίματος σε μια εκτεταμένη περιοχή (π.χ., μια θαλάσσια λεκάνη). Μέσω της χωρικής κατανομής διαφόρων στατιστικών μεγεθών (π.χ., μέση τιμή, μεταβλητότητα, συσχέτιση, τάση) και των παραμέτρων που χαρακτηρίζουν μια κατανομή είναι εφικτή η περιγραφή της μεταβλητής ενδιαφέροντος. Στο **Κεφάλαιο 2**, αναπτύσσεται το θεωρητικό υπόβαθρο της γραμμικής παλινδρόμησης, η οποία εξετάζει τη σχέση μεταξύ δύο ή περισσοτέρων εμπλεκόμενων μεταβλητών. Το μοντέλο γραμμικής παλινδρόμησης βασίζεται σε συγκεκριμένες παραδοχές (π.χ. ανεξαρτησία και κανονικότητα καταλοίπων) και οποιαδήποτε παραβίασή τους μπορεί να οδηγήσει σε εσφαλμένα συμπεράσματα. Από την άλλη πλευρά, τα ανεμολογικά και κυματικά δεδομένα συχνά παραβιάζουν ορισμένες από τις υποθέσεις της γραμμικής παλινδρόμησης λόγω σφαλμάτων, ανομοιογενών και άλλων παραγόντων επηρεάζοντας την εγκυρότητα των συμπερασμάτων (βλ. Παράρτημα Ε). Καθώς η εκτίμηση των παραμέτρων παλινδρόμησης με τη γνωστή μέθοδο ελαχίστων τετραγώνων επηρεάζεται από την ύπαρξη έκτροπων παρατηρήσεων, είναι σημαντικό να χρησιμοποιηθούν οι ανθεκτικές μέθοδοι παλινδρόμησης (*robust regression methods*) που θεωρούνται κατ' εξοχήν καταλληλότερες σε περιπτώσεις αποκλίσεων από τις υποθέσεις. Τέτοιου είδους ανάλυση κρίνεται αναγκαία στις περιπτώσεις π.χ., ύπαρξης πολλαπλών πηγών δεδομένων σε μια περιοχή με σκοπό την καλύτερη δυνατή διόρθωση των λιγότερο αξιόπιστων εξ αυτών. Στο τελευταίο κεφάλαιο του πρώτου μέρους (**Κεφάλαιο 3**), εξετάζεται η μοντελοποίηση ακραίων τιμών γραμμικών μεταβλητών λαμβάνοντας υπόψη και τη μεταβλητότητα των αντίστοιχων κατευθυντικών μεταβλητών. Σημειώνεται ότι η κατανόηση της συμπεριφοράς των ακραίων τιμών μετεωρολογικών και κυματικών δεδομένων συναρτήσει της κατευθυντικότητας είναι πολύ σημαντική κυρίως για το σχεδιασμό θαλάσσιων κατασκευών. Οι δύο επικρατέστερες θεωρήσεις για την εκτίμηση παραμέτρων ασυμπτωτικών κατανομών από δεδομένα είναι: i) η μέθοδος *block maxima*, στην οποία λαμβάνονται υπόψη οι μέγιστες τιμές διατεταγμένων τυχαίων μεταβλητών, και ii) η μέθοδος *peaks-over-threshold*, στην οποία εξετάζεται η ακολουθία ανεξάρτητων και ισόνομα κατανεμημένων τυχαίων μεταβλητών με τους όρους της να υπερβαίνουν ένα δεδομένο επίπεδο. Οι ανωτέρω μέθοδοι χρησιμοποιούνται στην ωκεάνια μηχανική για την πρόβλεψη (εκτίμηση) τιμών σχεδιάσεως και των αντίστοιχων περιόδων επανεμφάνισης της υπό μελέτη μεταβλητής. Στα πλαίσια αυτής της διατριβής, χρησιμοποιείται η δεύτερη θεώρηση, με χρήση της κατανομής Generalized Pareto, η οποία θεωρείται η πλέον κατάλληλη για τη συγκεκριμένη

περίπτωση. Στο πλαίσιο αυτό, οι παράμετροι της εκφράζονται συναρτήσει της κατευθυντικής μεταβλητής μέσω μιας σειράς Fourier. Για τον υπολογισμό των παραμέτρων της κατανομής Generalized Pareto προτείνεται η μέθοδος μέγιστης πιθανοφάνειας με την εισαγωγή ενός επιπλέον όρου ποινής που εξασφαλίζει σταθερότητα στα αποτελέσματα ανεξαρτήτως των όρων της σειράς Fourier. Επίσης, εξετάζονται διάφοροι μέθοδοι για τον υπολογισμό της τιμής κατωφλίου και την απο-ομαδοποίηση των δεδομένων, ενώ διερευνάται και η επίδραση διαφόρων συνδυασμών στο κατευθυντικό μοντέλο ακραίων τιμών. Τα αριθμητικά αποτελέσματα που προκύπτουν από την συστηματική μελέτη της ανωτέρω προσέγγισης σε ανεμολογικά και κυματικά δεδομένα στη Μεσόγειο Θάλασσα και εφαρμόζονται στους κλάδους της κλιματολογίας και των ΘΑΠΕ παρουσιάζονται στο δεύτερο μέρος (**Κεφάλαιο 4**).

Το τρίτο και τελευταίο μέρος (**Κεφάλαιο 5**) είναι αφιερωμένο στο παράκτιο περιβάλλον και αντιμετωπίζει ζητήματα διάβρωσης των ακτών σε αμμώδεις παραλίες λόγω της δράσης των κυμάτων. Για την επίτευξη των στόχων αυτού του μέρους, έγινε χρήση ενός ενδεδειγμένου πακέτου λογισμικού, του *MIKE 21/3Coupled Model* που αναπτύχθηκε από το *Danish Hydraulic Institute* (DHI). Το μοντέλο αυτό κάνει σύζευξη ενός φασματικού κυματικού μοντέλου για την παραγωγή των κυματογενών δυνάμεων, ενός υδροδυναμικού για την παραγωγή ρευμάτων και της στάθμης του νερού και ενός μοντέλου ιζηματομεταφοράς για την μεταβολή της βαθυμετρίας.. Δύο περιοχές, ευάλωτες σε φαινόμενα διάβρωσης, επιλέχθηκαν για τη μελέτη αυτή· η πρώτη είναι η παραλία της Βάρκιζας, στο Σαρωνικό κόλπο, με ένα έντονο χαρακτήρα διαμορφωμένο σε ένα αστικό περιβάλλον, και η δεύτερη είναι η ακτή της Σητείας, στην ανατολική Κρήτη, μια παραλία με άξονα ανάπτυξης τον τουρισμό. Και στις δύο περιπτώσεις, οι συνέπειες των υψηλών ρυθμών διάβρωσης θα ήταν επιζήμιες σε οικονομικό και κοινωνικό επίπεδο.

Η εργασία ολοκληρώνεται με τη διατύπωση των σημαντικότερων συμπερασμάτων της εργασίας και σχετικών προτάσεων για μελλοντική έρευνα πάνω στους άξονες που εξετάστηκαν (**Κεφάλαιο 6**).

Acknowledgments

Contents

List of Figures

List of Tables

List of Notation

The main notation used in this thesis are listed below. Since there is a large number of parameters, there are cases where the same symbol has been used twice. For such cases, the specific chapter or section is provided into parenthesis. All notations are also explained at the appropriate place in the text.

a, d, s (Sec. 1.2.1)	shape parameters of a linear distribution
b	scale parameter of a linear distribution
CV	Coefficient of Variation
c, g (Sec. 1.2.1)	location parameters of a linear distribution
d_{50}	median size of sediment particle
d_{ij}	normalized deviation
$d(\psi, \omega)$	circular distance between two angles ψ, ω
$F_X(x)$	cumulative distribution function of the linear random variable X
$F_\Theta(\theta)$	cumulative distribution function of the directional random variable Θ
$F_{X,\Theta}(x, \theta)$	joint (bivariate) cumulative distribution function of the random variables X, Θ
$f_X(x)$	probability density function of the linear random variable X
$f_\Theta(\theta)$	probability density function of the directional random variable Θ
$f_{X,\Theta}(x, \theta)$	joint (bivariate) probability density function of the random variables X, Θ
f_w	wave friction factor
g	gravitational acceleration
H_S (or H_{m_0})	significant wave height
h, h^*	bandwidth of kernel for linear and directional random variables
I_0	modified Bessel function of the first kind and zero order
IF	influence function of estimator
$K(x)$	kernel density function
Ku	Kurtosis
$L(\cdot)$	likelihood function for maximum likelihood method
$l(\cdot)$	negative log-likelihood function for maximum likelihood method
$l_p(\cdot)$	Penalized negative log-likelihood function for maximum likelihood method
$M(\theta)$	circular kernel density function
N (Sec 4.4.3)	total number of non-empty cells
N (Sec 5.2)	relative angular frequency
N_T	total number of cells
n	sample size
$N_{ij}^{(o)}$	observed number of data points falling in the $(i, j)-$cell
P	wind power density
$\overline{p}^{(o)}$	average observed frequency
$p_{ij}^{(e)}$	fraction of points from the estimated bivariate distribution that belong in the $(i, j)-$cell
$p_{ij}'^{(e)}$	difference between the estimated and the mean frequency
$p_{ij}^{(o)}$	fraction of points from a data set that belong in the $(i, j)-$cell
$p_{ij}'^{(o)}$	difference between the observed and the mean frequency
q (Sec. 1.1, 4.4)	number of parameters of a particular distribution
q (Sec. 5.5.2)	seabed level change
q_B, q_S	bed and suspended load transport

Symbol	Description
$R^2_{a,1}$	coefficient of determination for the univariate case
$R^2_{a,2}$	coefficient of determination for the bivariate case
r_{FGM}	statistical association parameter between X and Θ in Farlie-Gumbel-Morgenstern model
SC	sensitivity curve of estimator
Sk	Skewness
s (Ch. 4)	Standard deviation
T_{LR}	likelihood-ratio test statistic
T_P, T_z	peak and zero up-crossing wave period
u	threshold for peaks-over-threshold method
u_{bw}	bottom (or near-bed) wave orbital velocity.
w (Sec. 3.7.1)	weight of penalty term
$w_i(\cdot)$ (Sec. 2.4.4)	weight function for generalized M – estimator
X, Y	linear random variables
$z^{(\alpha)}$	100α−th percentile point of a standard normal distribution
β_0, β_1	intercept and slope in linear regression model
β_0', β_1'	intercept and slope in inverse regression model
γ^*	gross error sensitivity of estimator
δ, ε	error terms in regression
H, Ξ	true (latent) random variables in regression such that $H = \beta_0 + \beta_1 \Xi$
ϑ	set of unknown parameters
θ^*, θ_{cr}	dimensionless and critical Shields parameter
Θ, Ψ	directional random variables
κ, ρ (Sec. 1.2.2)	concentration parameters of a circular distribution
μ (Sec. 1.2.2, 3.3)	location parameter of a distribution
ξ	shape parameter for extreme value distributions
ρ (Sec. 1.3.1)	Pearson product-moment correlation
ρ, ρ_s (Sec. 5.3.2)	density of fluid, density of sediment
$\rho(\cdot)$ (Sec. 1.3.1)	objective function for M – estimator
$\sigma_{ij}^{(e)}$	expected standard deviation for the number of data points falling in the (i,j) −cell according to the Binomial distribution
τ_{bc}, τ_{bw}	current- and wave-produces bed shear stress
τ_{cr}	critical bed shear stress
$\Phi(\cdot)$	cumulative distribution function of the standard Normal distribution
χ^2	chi square test or chi square statistic
ψ_P	'correlation-type' parameter in Plackett model
ψ (Sec. 1.3)	parameter representing the dependence structure between X and Θ in Johnson-Wehrly model
$\psi(\cdot)$ (Sec. 2.4.4)	score function for M – estimator
ω	weighting factor for a mixture distribution

List of Abbreviations/Acronyms

A-D	Andreson–Darling goodness-of-fit test
AWS	Archimedes Wave Swing
BIC	Bayesian information criterion
BLUE	Best Linear Unbiased Estimate
cdf	cumulative distribution function
CV	Coefficient of Variation
DeCA	DeClustering Algorithm
DHI	Danish Hydraulic Institute
FGM	Farlie-Gumbel-Morgenstern bivariate model
GAM	Gamma distribution
GEV	Generalized Extreme Value distribution
GP	Generalized Pareto distribution
GW	Gamma-Weibull mixture distribution
HD	HydroDynamic (module)
IAV	Inter-Annual Variability
JSB	Johnson S_B distribution
JW	Johnson-Wehrly bivariate model
ISE	Integrated squared error
K-S	Kolmogorov–Smirnov goodness-of-fit test
KAP	Kappa distribution
kdf	kernel density function
LAD	Least Absolute Deviation estimator
LGN	Lognormal distribution
LMS	Least Median of Squares estimator
LR	Likelihood-Ratio test
LTS	Least Trimmed Squares estimator
MISE	Mean integrated squared error
ML	Maximum Likelihood
MAV	Mean Annual Variability
MCAE	Mean Circular Absolute Error
MEM	Measurement Error Model
MRB	Mean Relative Bias
mvM	mixture of von Mises distributions
NN	Truncated Normal mixture distribution
NWP	Numerical Weather Prediction
OWF	Offshore Wind Farm
OLS	Ordinary Least Squares estimator
OWC	Oscillating Water Column
OWSC	Oscillating Wave Surge Converter
pdf	probability density fnction
PLA	Plackett bivariate model
POT	Peaks-Over-Threshold method
RME	Root Mean Error
RMSE	Root Mean Square Error
rv	random variable
SI	Scatter Index
SRS_E	Sum of relative squared error
SS_E	Sum of squared error
SS_T	Total sum of squares
ST	Sediment Transport (module)
SW	Spectral Wave (module)
vM	von Mises distribution
WAK	Wakeby distribution

WC	Wrapped Cauchy distribution
WEC	Wave Energy Converter
WEI	Weibull distribution
WGEV	Weibull-Generalized Extreme Value mixture distribution
WN	Wrapped Normal distribution
WW	Weibull mixture distribution

Introduction

Background and research motivation

Metocean (i.e. meteorological and oceanographic) variables play a critical role in a variety of highly interrelated physical processes encountered in the marine environment. Among the multitude of variables, wind and wave parameters control the air-sea exchange of energy mass and momentum and, by extension, various scientific fields, with both socio-economic and environmental impacts, are influenced. In this context, the accurate knowledge of the most important aspects of wind and wave climate and the reliable estimation of extreme events so as to contribute in the reduction of risks, are of fundamental importance. Closely related areas of application of climatological research are engineering projects such as the design and construction of coastal infrastructures, offshore activities (e.g. platforms for oil extraction), air and water pollution dispersion, ship routing and scheduling, sediment transport and coastal erosion/accretion, and marine renewable energy sector with an increasing interest for development the last two decades.

In general, the metocean climate system cannot be described in full detail since its components are controlled by physical laws and countless factors, either adequately or partially known, that induce instabilities and nonlinearities. Therefore, the need of probabilistic concepts and statistical tools is inevitable in order to describe the metocean climate at a particular location and time period. The most typical linear random variables used to characterize wave and wind conditions are significant wave height and mean wave period, and wind speed, respectively. However, the description of the wave and wind climate is limited if these parameters are solely provided; thus, a more integrated assessment is needed in order to define the climate accurately in an area of interest. In this context, the corresponding directional variables (i.e. mean wave and wind direction) complete the description of the local/regional wave and wind climate and should always be incorporated in such analysis. The importance of including directionality has been highlighted in many previous studies concerning, for instance, marine renewable energy (Porté-Agel et al., 2013; Atan et al., 2016; Hildebrandt et al., 2019) and coastal erosion (Harley et al., 2017; Mortlock et al., 2017; Yanalagaran and Ramli, 2018).

Analysing directional data, one of the main aspects in this thesis, is a rather old subject in mathematical statistics; however, the advance of this field follows a slow pace compared to the statistical analysis of linear data. It can be said that the 1900s was a milestone for the initiation of modern directional statistics with Rayleigh (1880; 1905; 1919), Kluyver (1906) and Pearson (1905a; 1905b; 1906) studying the uniform random walk on the sphere, and von Mises (1918) introducing a distribution on the circle. Yet, after 1953, Fisher, Watson, Mardia, Batschelet and other researchers contributed to the essential growth of studying directional data, taking into account the curvature of the sample space, via numerous techniques and directional distributions. Directional observations are mainly regarded as points (or vectors) either on the circumference of the unit circle of $\mathbb{R}^2$, or on the surface of the unit hypersphere of $\mathbb{R}^d$, $d \geq 3$, referred to as circular and spherical data, respectively. Diverse scientific disciplines deal with measurements that are recorded as angles, orientations or directions. Typical examples of directional data in physical and life sciences include wind direction measurements in meteorology, wave and current directions in oceanography, the study of directions (orientations) of birds (animals) and movements of organisms in biology and ecology, respectively, the analysis of geological phenomena such as the orientation of rock fractures and cores in geology, the determination of the location of an epicentre of an earthquake in earth sciences, and the description of the motion of celestial bodies in astronomy. The sense of rotation (clockwise, counter clockwise), the arbitrary choice of a unique origin (called zero direction), the lack of notion of minimum and maximum values (ranking), the coincidence of

the "start" and "end" of their range (i.e. $0 = 2\pi$), and the inherent periodicity (i.e. two points θ and $\theta + 2k\pi$, $k = 0, \pm 1, \pm 2, ...$, represent the same point on the unit circle) are some distinctive characteristics of circular data that render their statistical analysis dissimilar from linear data; thus, the commonly used statistical tools applied for the dominant linear variables, from the calculation of simple descriptive statistics to statistical inference, are not appropriate.

Marine renewable energy and coastal erosion, the two main pillars of application throughout this thesis, come under the umbrella of climate change. Climate change mitigation requires changes in the global energy system with oceans offering a vast source of renewable energy that up to now has not been utilised on a significant level despite its great potential. In order to meet the EU targets by 2030 and 2050 as regards energy generation from renewables, the share of marine renewable energy in the final energy consumption should be increased. On the other hand, the reliable operation, financial viability and local environmental impacts of structures and devices deployed in the marine environment require the accurate knowledge of metocean climatology and climate variability, and examination of their response under extreme conditions. Although nowadays there are plenty of data sources for metocean variables with a reasonable spatial and temporal coverage and resolution, the need to reduce uncertainties and improve our knowledge of the marine environment is vital so that the associated risks can be quantified. Moreover, climate charge is intensifying the problem of coastal erosion, a global problem that is threatening human activities (e.g. tourism, commerce), properties and infrastructures along with the biotic and abiotic elements of the coastal environment. The understanding of sediment transport mechanism in the coastal environment is of paramount importance for the accurate prediction of shoreline evolution and seabed changes, a rather challenging issue due to the highly complex processes involved. These processes are under perpetual changes, which greatly vary in duration and geographic scale, towards an equilibrium state. The use of numerical models is widely implemented for the simulation and modelling of sediment transport with the possibility to consider different wave conditions and bathymetric scenarios and obtain results in a reasonable time frame; see the exhaustive review of Papanicolaou et al. (2008).

Research aims and objectives

The present thesis attempts to provide a holistic approach for the probabilistic modelling of metocean linear and directional variables. The mathematical tools and methodologies presented do not provide an exhaustive means for the metocean climate description and modelling but focuses on particular aspects that are either not known or less applied. Furthermore, the analysis can be easily extended to other relevant metocean variables, such as current, tidal elevation, air and sea temperature and density, salinity and solar radiation, and other fields of geosciences as well, such as meteorology, geology, geohazards (e.g. earthquakes, floods), geography and ecology.

Depending on the objectives and requirements of the application, the spatial scale of interest should be at first defined in order to acquire the most suitable, in terms of spatiotemporal coverage, data set with the corresponding time series of the linear and directional variables for the analysis. Even though large-scale studies (e.g. at the global level) can provide a reliable indication of general long-term trends, they are technically unable to resolve the spatial variability in a local wind and wave climate scale. In this respect, long-term measurements obtained by *in situ* devices, such as meteorological masts or lidars and oceanographic buoys, are preferable for local assessment purposes since they are considered to be the most reliable, in terms of quality, data sources, although they usually have a scant temporal extent. On the other hand, when a larger region is of interest (e.g. a sea basin, world ocean), the most appropriate data sources for deriving climatologies are gridded data coming either from satellite observations and other remote sensing instruments (e.g. radar altimeters, scatterometers) or

hindcast data, i.e. reanalysis of meteorological data and large-scale atmospheric models producing time series of metocean parameters extending back in time many years. In this context, the assessment is usually carried out on an annual, seasonal or monthly basis by providing the corresponding low-order statistical characteristics of the examined variables, such as mean values and variances, along with additional statistical parameters that quantify the corresponding temporal variability; see also Figure 0-1. Apart from the identification of the climate structure for these temporal scales in an area, the analysis of the longer-term changes and relevant variability is also important in real-life situations such as ocean energy economics; hence, identification of changes in the multiannual (e.g. decadal) scale can also be examined if the duration of the available data set permits it. Regardless of the spatial scale, efficient and reliable metocean climate modelling and estimation of the corresponding extremes requires at least 30 years of time series (World Meteorological Organization, 2017).

Aim 1. Regression/calibration of linear and directional variables
However, the gridded/simulated data suffer from various types and degree of uncertainties and measurement errors; see Table E-6-1. The horizontal resolution of gridded data is also an important parameter, mainly for nearshore areas concerning wind and wave climate studies. Let us note that the accurate wind modelling in coastal areas remains an open issue due to the influence of many complex factors, such as land/sea distribution and corresponding thermal and roughness gradients and local topography. On the other hand, regarding sea waves, the intense spatial variations of the coast and bottom depth along with the various mechanisms that take place in the coastal zone require high-resolution grids to resolve the small-scale changes. Therefore, there is need to use reliable data for validation purposes of the gridded data in case they are available in the examined region (Menendez et al., 2014). Based on the knowledge of a physical process, relationship functions among involved variables (e.g. modelled and measured data) can be derived by means of statistical methods with the most common one the regression model. Based on the regression models and in case of multiple data sources in an area, (linear and circular) calibration methods are essential to be applied in order to correct as much as possible the less accurate data sources. The linear regression model demands the fulfilment of certain statistical assumptions (e.g. independent and normally distributed errors), since any violation of these assumptions may lead to erroneous results and invalid conclusions. Metocean data may often violate some of the assumptions of the linear regression model due to data errors and inhomogeneities and other factors, and affect the validity of the prediction or inference. Since regression estimates obtained by the method of least squares are affected by the presence of outliers, it is important to detect through diagnostic tools potential outliers. Another way to deal with outliers and small deviations from the assumptions is the use of robust estimators, which can still deliver results of sufficient accuracy. It is also important to mention that prior to the use of a data set, the obtained data should be quality checked and validated in order to confirm that the data set is correct and can be used for the purposes of the application.

Aim 2: Probabilistic modelling and analysis of linear and directional variables
Uni- and multivariate probabilistic models are applied to the obtained dataset for the description of the assumed population and the accurate estimation of the statistical characteristics of the metocean variables. The optimum choice of the model is based on some goodness-of-fit criteria. In this way, it is possible to assess the variability of an energy conversion system at a specific site and minimize uncertainties in resource estimates during the phase of planning. The joint probability analysis, used primarily to make predictions about probabilities of occurrence of specific sea-states and wind conditions for dependent random variables, can improve the accuracy of the results. Probabilistic modelling of metocean data can be implemented through parametric and non-parametric models that permit the characterization of the wind and wave climate under different assumptions, both presenting advantages and shortcomings.

When a larger area is of interest, the results of the climate analysis are provided as a spatial distribution through standard statistical measures and tools. Such tools include sample

descriptive statistics, correlations, trends and significance tests while the parameters of any parametric model, either concerning the entire data sample or extreme data, can be also presented in the spatial scale. Apart from the standard statistical parameters (mean value, variance), the mean annual and inter-annual variability can be provided, along with dependence structures of linear, directional and linear-directional variables. Moreover, the quantification of climate changes can be provided through trend estimation, where parametric and non-parametric models can be employed for linear random variables. As regards changes of the directional variables, the estimation of a linear trend (slope) for directional time series is not feasible. Hence, the angular distance can be provided for different temporal scales (e.g. from year to year). Significance testing is necessary to assess whether the obtained trends do exist and are not a result of strong (e.g. seasonal) variations. Statistically significant trends can be identified via parametric and non-parametric tests, with the latter being more robust in the presence of outliers.

Aim 3: Directional extreme value analysis

Furthermore, the understanding of the behaviour of extreme values coming from metocean data is of paramount importance particularly for the design of marine structures. In order to select the extremes from the available data sample and fit them with the corresponding extreme value distribution model, two approaches can be implemented: i) block maxima approach, where a set of maxima (or minima) of the variable is identified by the whole data set and is modelled by the Generalized Extreme Value distribution, and ii) the peaks-over-threshold approach, where the extreme values are selected over a predefined level and the Generalized Pareto distribution is used to fit this extreme value data set. Using one of the above distributions, the next step is statistical inference on the data; thus, the return levels, associated with certain return periods, can be estimated through the inverse distribution function. However, there are cases where the assumption of statistic homogeneity is violated because of directional (and/or spatial and temporal) variations, affecting in turn the wind and wave regimes. In this respect, modelling of linear metocean variables relies on the variability of directional ones.

Summarizing the above discussion on metocean climate analysis and modelling, the following objectives for the present dissertation were set:

- to develop an integrated approach for climate modelling of linear and directional metocean random variables;
- to highlight new features regarding directional climate variability, which has received less attention;
- to study and compare consistently various parametric models for linear and directional variables in the univariate and bivariate case through statistical metrics;
- to evaluate thoroughly parametric and non-parametric models for the joint description of linear and directional metocean variables;
- to provide a robust methodology for the assessment of less reliable data sources, propose methods that take into account outlying observations for linear variables, and correct directional metocean variables, which is rather uncommon, by identifying the most appropriate model;
- to use extreme value analysis methods that include directionality, introduce a new penalised likelihood criterion for parameter estimation, and investigate its behaviour under different methods of threshold selection and declustering.
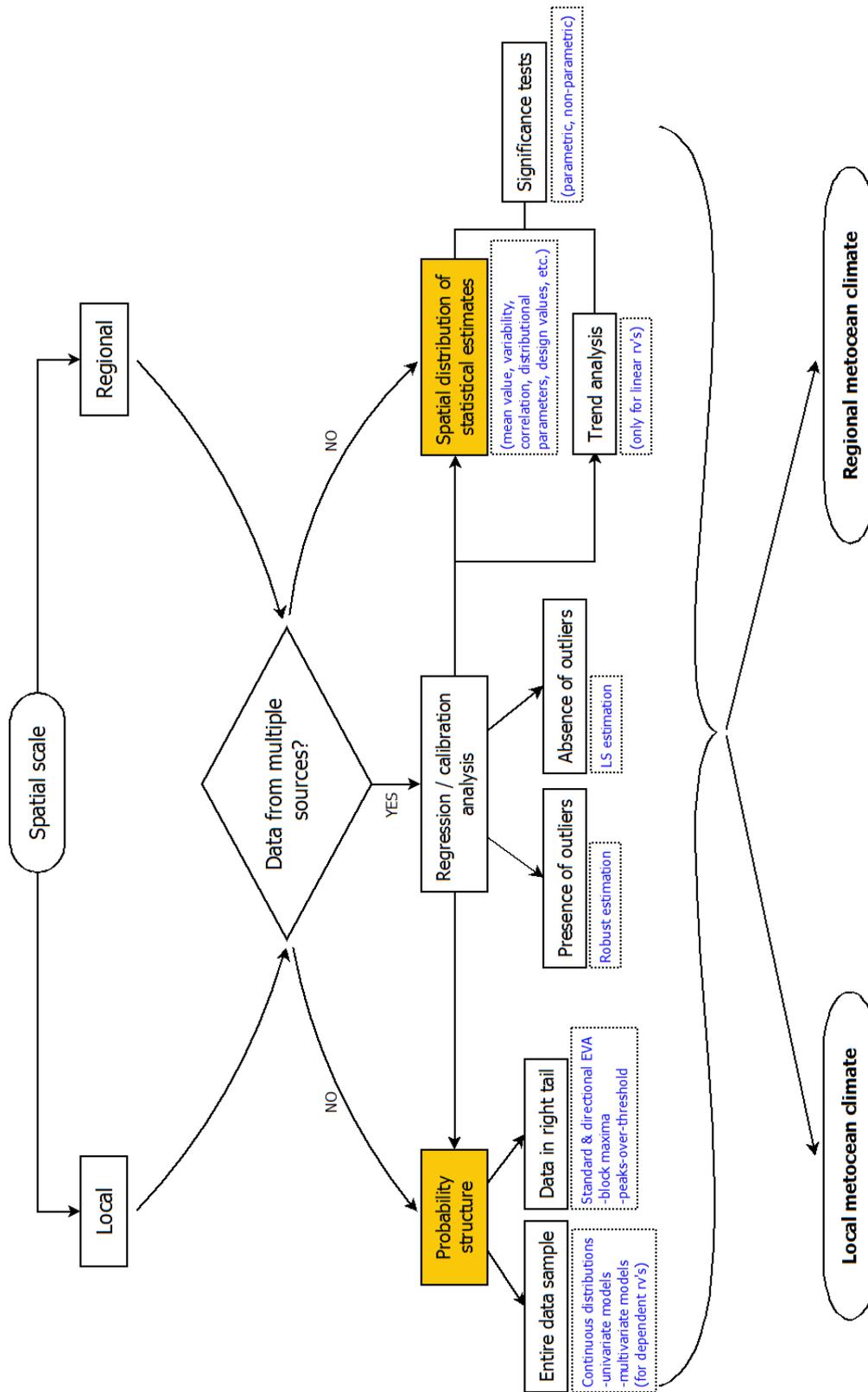
Figure 0-1. Diagram for the representation of the metocean climate analysis within the scopes of this thesis.

Aim 4: Impacts of wave action on sandy beaches

Estimation and prediction of sediment transport patterns and seabed changes due to wave action is the other main goal of this thesis. The varying time scales and amplitude of waves, from hours regarding storm waves to seasons regarding the typical wave climate, result in a variable behaviour of the sediments in the coastal zone. For this reason, sediment transport is examined

i) under the action of individual high waves for a short time window, and ii) under the cumulative action of waves over time (e.g. throughout a typical year). Let us note that the above-mentioned mathematical/statistical tools can be implemented in the context of sediment transport modelling, such as the calibration methods for linear and circular variables.

In this connection, some additional objectives for the present dissertation were set:

- to assess the impacts of waves and circulation on sediment transport processes during and after intense sea states;
- to provide a cost-efficient and replicable method, within the philosophy of wave input reduction techniques, for the estimation and prediction of seabed level by considering the accumulative action of waves.

For the above purposes, a well-established and integrated software was used, the MIKE 21/3 Coupled Model developed by the Danish Hydraulic Institute (DHI). The applied modelling system couples internally a spectral wave model with a hydrodynamic model and a sediment transport model due to the interdependence of the involved complex physical processes. In brief, the wave-induced forces computed by a wave model are provided to a hydrodynamic model to calculate wave-related phenomena (e.g. wave-induced currents) by considering additional processes as well, such as wind forcing. On the other hand, the hydrodynamic model returns water levels and currents to the wave model. Then, the sediment transport process is controlled by the bed shear stress, which is induced by waves and currents. The total sediment transport alters the bathymetry, which in turn, affects the wave and current fields. In this regard, two real coastal sites, vulnerable to erosion phenomena, were selected during this research; the first one is Varkiza beach, located in the Saronic Gulf, with an intense recreational character in an urban environment and the other one is Sitia beach, situated in the eastern part of Crete Island, a typical tourism-oriented beach. The consequences of high coastal erosion rates for both of them would be detrimental for the local societies and economies.

Both objectives can contribute to understand better the dynamics of a coastal system, identify erosion/accretion patterns and predict quickly and efficiently potential future trends, which can be valuable for planning and managing coastal activities. Let us remark that it is not expected to predict with absolute accuracy the values of seabed level because of the high level of inherent uncertainties but rather evaluate the skill of the model as regards the relevant trends compared with observations. As was aptly expressed by Klonaris et al. (2018) as regards the tools for coastal sediment transport and geomorphology: "*The complexity and uncertainty of the various processes is so intense that predictions of sediment loads within a factor of 2, or even 5, are generally considered as satisfactory, especially for field measurements*". Besides, it has been shown that sediment transport formulas may provide good results when compared with laboratory experiments (since they rely more on laboratory than field data for validation/calibration purposes) but they have an inadequate performance for real-scale conditions (Li and Huang, 2013) while numerical models provide poor predictions when they are not properly calibrated (Do et al., 2018).

Innovative contributions and original publications

Along the research of the above mentioned aims, the following original contributions that worth mentioning have been achieved:

- Structure of an integrated approach for climate modelling of linear and directional metocean random variables. In this context, new features regarding directional metocean climate variability are highlighted, which are investigated along with the corresponding linear variable(s).

- Thorough and quantitatively consistent evaluation of parametric and non-parametric models for the bivariate case. This analysis revealed that the models that have the best performance in the univariate case do not ensure their performance when the joint description of linear and directional variables is considered. In addition, the bivariate parametric model of Johnson and Wehrly (between linear and directional variables) is recommended to be also examined even if non-parametric models are considered.

- A methodology is proposed to evaluate thoroughly regression/calibration models for the correction of linear and directional variables obtained from less reliable data sources. For a reliable assessment, a reference data source is required, ideally measurements from metocean devices, and concurrent data samples of one-year duration at least. Specific regression/calibration models are proposed for both linear and directional metocean variables.

- A new statistical metric for the evaluation of circular calibration models is introduced, named root mean error. It is dimensionless and the lower its value, the better the performance of the model.

- A penalised likelihood criterion is introduced for a more stable optimization of the estimated parameters of the directional extreme value model based on Generalized Pareto distribution. These parameters are expressed by means of a Fourier series expansion and even for higher-order expressions the solution is stable. A thorough analysis is also performed as regards various methods of threshold selection and declustering in order to obtain a better understanding of the effects of their different combinations on the estimation of the Generalized Pareto parameters and the design values of linear variables taking into account directionality effects.

- A new method is developed based on wave input reduction techniques for the estimation of seabed level in order to reduce the computational time of morphological simulations. Apart from some basic hydrodynamic parameters and sediment characteristics, the joint probabilistic behaviour of wave height and wave period, which are usually available from numerical models, is examined and the corresponding combinations that contribute to the initiation of sediment movement are identified. This approach is implemented at a sandy beach and the obtained results are compared against the results from the full range of wave conditions and the situation encountered in reality.

Parts II and III contain various articles about a specific topic that is related with the description of metocean climate with applications in the ocean and coastal environment, respectively. Therefore, each of the sections of Part II and the last two sections of part III is connected with one paper, and in one case with 3 papers, where part of the results are presented.

As regards Part II, the following publications are considered along with the contributions of the author:

P.1. Soukissian, T., Karathanasi, F., Axaopoulos, P., Voukouvalas, E.G., Kotroni, V., 2018. **Offshore wind climate analysis and variability in the Mediterranean Sea**. *International Journal of Climatology* 38: 384–402.
The author contributed to the development of the statistical analysis and the visualization of the results, performed the statistical tests, investigated the directional changes, and contributed to the writing of most sections of the original draft, and the review and editing of the whole manuscript.

P.2. Soukissian, T.H., Karathanasi, F.E., 2016. **On the use of robust regression methods in wind speed assessment**. *Renewable Energy* 99: 1287–1298.
The author made all the statistical analysis, wrote all the sections, and contributed to the review and editing of the whole manuscript.

P.3. Soukissian, T., Karathanasi, F. Voukouvalas, E., 2014. **Effect of outliers in wind speed assessment**. Proceedings of the 24th *International Offshore (Ocean) and Polar Engineering* Conference, 1: 362-369, Busan, June 15–20.
The author made all the statistical analysis, wrote all the sections, and contributed to the review and editing of the whole manuscript.

P.4. Karathanasi, F.E., Soukissian, T.H., Axaopoulos, P.G., 2016. **Calibration of wind directions in the Mediterranean Sea**. Proceedings of the *26th International Ocean and Polar Engineering Conference*, 1: 491-497, Rhodes, Greece, June 26–July 1.
The author developed the study work plan, made all the statistical analysis and visualization of the results, wrote all the sections, and contributed to the review and editing of the whole manuscript.

P.5. Soukissian, T.H., Karathanasi, F.E., 2017. **On the selection of bivariate parametric models for wind data**. *Applied Energy* 188: 280–304.
The author contributed to the statistical analysis, wrote all the sections, prepared the visualization of the results, and contributed to the review and editing of the whole manuscript.

P.6. Karathanasi, F., Soukissian, T., Belibassakis, K., **Directional extreme value models in wave energy applications**. *Atmosphere*, in press.
The author selected the examined locations and obtained the wave data, made all the extreme value and statistical analysis, investigated the penalized maximum likelihood and the methods of threshold selection and declustering, made the visualization of the results, wrote all the sections, and contributed to the review and editing of the whole manuscript.

Paper P.1 is related to Section 4.2, P.2–P.4 to Section 4.3, P.5 to Section 4.4 and P.6 to Section 4.5.

As regards Part III, the following publications are considered along with the contributions of the author:

P.7. Belibassakis, K., Karathanasi, F., 2017. **Modelling nearshore hydrodynamics and circulation under the impact of high waves at the coast of Varkiza in Saronic-Athens Gulf**. *Oceanologia* 59(3): 350–36.
The author set up the coupled model, validated the model results against in situ measurements, made the visualization of the results, wrote all the sections, and contributed to the review and editing of the whole manuscript.

P.8. Karathanasi, F., Belibassakis, K., 2019. **A cost-effective method for estimating long-term effects of waves on beach erosion with application to Sitia bay, Crete**. *Oceanologia* 61(2):276–290.
The author contributed to the development of the proposed methodology, set up the coupled model, made the visualization of the results, wrote all the sections, and contributed to the review and editing of the whole manuscript.

Papers P.7 and P.8 are connected with Sections 5.4 and 5.5, respectively.

Thesis outline

The thesis is divided into three major parts. In the first part, comprised of three chapters, the theoretical framework for modelling linear and directional variables is developed. Results from real data sets stemming from the disciplines of climate modelling and marine renewable energy, most of which have been published or submitted for review, are presented in the second part. The third part is devoted to the coastal environment and addresses coastal erosion issues on sandy beaches due to the wave action.

In **Chapter 1**, the probabilistic modelling of metocean (linear and directional) variables is introduced. Parametric and non-parametric models are presented for the description of both single linear and directional variables (univariate case) and their joint behaviour (bivariate case). The parametric bivariate models are based on the marginal distributions of the corresponding linear and directional variables along with a dependence structure and the linear-directional density is estimated by the use of copulas, with Farlie-Gumbel-Morgenstern (1956, 1960) and Plackett (1965) families, and the model proposed by Johnson & Wehrly (1978). Kernel density estimators are used for the non-parametric univariate and bivariate modelling of both linear and directional variables.

Chapter 2 focuses on the calibration problem of linear and directional variables in regression analysis. Along with the simple linear model, more efficient estimators, the so-called robust estimators, are presented that are less sensitive in the presence of outlying observations, which result in violations from the standard assumptions of the former model. Circular regression models, which involve both response and predictor circular variables, are introduced along with calibration models, that have received lesser attention in the relevant literature. The latter are based on distance-based estimators according to the mapping of SenGupta et al. (2013).

Chapter 3 begins with the necessity of describing the extreme behaviour of wind and wave features in terms of directionality in the ocean energy technology sector. In this context, the foundations of the classical extreme value theory are presented and derivations of the peaks-over-threshold method are mentioned. Additionally, threshold selection and declustering methods are reviewed as their performance is assessed in the next chapter. The directional extreme value model, as proposed by Jonathan and Ewans (2007), is also presented and some comments are provided as regards the method of estimating the unknown parameters. A new rational for the parameter estimation is recommended based on a penalised likelihood criterion, which seems to be numerically stable for optimization, while a variety of methods as regards threshold selection and declustering are considered to examine their effects on the performance of the directional extreme value model.

In **Chapter 4**, numerical results from each of the above research directions are given in order to illustrate the performance of the proposed methods and tools from different metocean data sources. The first area of application deals with a thorough wind climate analysis and variability by means of statistical tools; new features regarding especially wind climate variability are highlighted. The second area is devoted to applications related with ocean energy assessment. Parametric univariate and bivariate models are applied to wind speed and direction data and the detailed evaluation leads to interesting findings as regards their performance. Parametric and non-parametric bivariate models are also evaluated for wave energy flux and wave direction, two parameters of high importance for the emerging wave energy sector. Results from the calibration of wind speed and direction obtained from different wind data sources are also presented by considering *in situ* measurements the reference source. In this section, a methodology is proposed for the assessment of less reliable data sources while a new statistical metric is introduced in the calibration of circular variables. The last section of this chapter concerns the application of the directional extreme value model, presented in the previous chapter, on wave data. Various numerical results are implemented in order to examine the effects of this model under the consideration of different methods as regards threshold selection and declustering techniques.

In **Chapter 5**, the coastal environment is examined by means of modelling wave action and sediment transport through a dynamically coupled modelling system. Coastal erosion attributed to wind-generated waves is examined under two perspectives. The first one considers storm events acting for a short time window at a sandy beach, where a plethora of measurements was available, that used not only to validate model results but for qualitative comparison purposes of seabed level change as well. The second point of view takes into account the entire wave

action throughout a typical year and introduces a cost-effective methodology, following the rationale of wave input reduction techniques, for the estimation of seabed level. In this respect, the wave conditions of the full time series are reduced to some representative conditions based on the Shields criterion, used as a determinant for the initiation of sediment movement. The results from this technique are compared against the full range of wave conditions and a parallelism is made with the real situation encountered at the examined area.

Chapter 6 recapitulates the most important results of this thesis and provides further directions for future research.

Finally, in the **Annexes**, some supplementary concepts and tools are provided. For the sake of completeness, the descriptive statistics of directional variables are summarized (Appendix A), various statistical tools (e.g. correlation measures, evaluation metrics) implemented in metocean climate modelling studies are determined (Appendix B), the maximum likelihood (ML) method for the estimation of the Generalized Pareto distribution is shortly presented (Appendix C), the basic mathematical formulation of the numerical modelling package used for the purposes of this thesis, MIKE 21/3 Coupled Model Flexible Mesh developed by the Danish Hydraulic Institute, is defined (Appendix D), and the main categories of metocean data sources along with the particular datasets that are analysed in Chapter 4 are provided (Appendix E).

PART I
PROBABILISTIC APPROACHES FOR MODELLING LINEAR & DIRECTIONAL VARIABLES

Chapter 1 Construction of probability distributions of linear and directional variables

1.1 General

Metocean conditions have clearly a significant influence on a variety of random physical processes that take place in the marine environment and interact with offshore/nearshore facilities and coastal infrastructure. Appropriate probability models of metocean variables at a location of interest are evidently a useful tool for the estimation of the corresponding conditions and accurate quantification of their frequency of occurrence. The existence of long-term data are necessary for the probabilistic modelling of metocean variables, which can be obtained by measuring devices, remote sensors and numerical simulations (usually hindcasts).

A special application of long-term modelling of metocean variables refers to marine renewable energy and coastal morphology, which are the main application areas in the present thesis. Metocean variables (e.g. winds, waves, currents) are required for each lifecycle phase of a marine energy project, from planning to decommissioning, in order to ensure safety and reliability of the design of the structure and high performance during operation, while wave and winds dominate in coastal processes and determine to a great extent beach morphology.

In this thesis, the analysis of the metocean variables is confined to linear variables of wind conditions and sea states, including wind speed, significant wave height, and wave period (energy or peak period) along with the corresponding directional features, i.e. wind direction and wave direction, which are of great interest in ocean and coastal engineering applications. The probability distributions of the available metocean data offer an essential understanding of their characteristics and features, from which several descriptive statistics can be derived to summarize the bulk statistical properties of metocean variables at different locations, and their use is inevitable due to the randomness of the involved phenomena and the lack of thorough knowledge from the data. Since such models cover the entire range of values of the examined variable (both body and tail regions), they are preferred in cases a system depends on mean, calm and storm conditions (e.g. beach response to wave forcing).

Joint probability distributions of different metocean parameters have received increasing attention in order to perform a more realistic analysis due to the inherent complexity of the above systems, facilitated by the recent development of numerical models and the direct availability of long-term metocean data. Although several research studies have examined joint statistical models for metocean variables and different approaches have been recommended for the estimation of bivariate probability distributions, there seems to be no general agreement yet. For instance, Vanem (2016) concluded that "*multivariate modelling of met-ocean conditions remains a challenge, even in the bivariate case*" in the context of presenting bivariate models for wave data.

This chapter deals with the probabilistic modelling of the abovementioned metocean variables by means of univariate (parametric and non-parametric) models for both linear and directional variables and bivariate (parametric and non-parametric) models for the joint description of linear and directional variables, based on the marginal distribution of the corresponding variables along with a dependence structure. The main motivations for working on the theoretical aspects of this subject are the following: from a recent study (see Section 4.4), it has been revealed that there is inconsistency of univariate models when the joint description between two variables is examined. Specifically, it will be shown in the applications of Part II that the best univariate models of linear variables do not ensure that will also provide the best fits when considering the joint description of linear and directional variables. Moreover, there

are cases where the parametric bivariate models outperform the non-parametric ones under various statistical measures, which gives rise to a more thorough examination when probabilistic models are sought for such problems. This justification is strengthened by the better statistical properties of the parametric models, their facilitation in making predictions and easier estimation of the unknown parameters.

1.2 Univariate case

For the accurate description of metocean characteristics, the determination of the corresponding probability density function (pdf) is essential in renewable energy applications. For instance, the average power of a wind turbine is directly associated with the pdf of wind speed u_w and the corresponding wind turbine power curve $P_w(u_w)$, obtained by $\int_0^\infty f(u_w) P_w(u_w) du_w$. Given that the latter parameter is known rather accurately by the wind turbine manufacturer, the significance of modelling wind speed as precise as possible is imperative in order to minimize wind power estimation errors in the phase of wind resource assessment (Morrissey et al., 2010; Ouarda et al., 2015; Rodriguez et al., 2015) and reach safer decisions on wind turbine selection and economic evaluation of a wind farm.

In the relevant literature, wind speed is traditionally modelled as a two-parameter Weibull distribution mainly due to its simplicity and flexibility; see, e.g. Fyrippis et al. (2010); Rocha et al. (2012); Arslan et al. (2014). However, as was noted by Chang (2011), there are many limitations, e.g. inefficiency in the accurate modelling of both calm winds (Drobinski et al., 2015) and wind speeds higher than 14 m/s (Sarkar et al., 2017), while its use has not been justified for modelling wind speed data; see also Jourdier and Drobinski (2017). Alternative distributions have been proposed in a number of studies that examine a great variety of coastal and offshore wind speed regimes, proving their superiority and thus providing better options for fitting wind speed. Among them, there are: i) the conventional unimodal distributions with three or more parameters that include Kappa (Ouarda et al., 2015; Ouarda et al., 2016), Wakeby (Morgan et al., 2011), Johnson S$_B$ (Soukissian, 2013), three-parameter Weibull (Stewart and Essenwanger, 1978), Gamma (Dong et al., 2013), Lognormal (Alavi et al., 2016a; Alavi et al., 2016b) and Nakagami (Alavi et al., 2016a; Dookie et al., 2018) distributions; ii) the multimodal parametric distributions, which are used to adequately represent wind regimes with particularities, such as the mixture distributions including the Weibull mixture (Carta and Ramirez, 2007; Akpinar and Kavak Akpinar, 2009; Qin et al., 2012), the Weibull- and Lognormal-Generalized Extreme Value mixtures (Kollu et al., 2012), the Normal mixture (Chang, 2011), the Gamma mixture (Ouarda et al., 2015) while a variety of heterogeneous mixture distributions has been assessed in Shin et al. (2016) and Ouarda and Charron (2018), and the Maximum Entropy type distributions (Chellali et al., 2012; Zhang et al., 2014), and iii) the non-parametric distributions (Jeon and Taylor, 2012; Zhang et al., 2013; Hu et al., 2016); for instance, kernel density estimation is one of the most powerful techniques in terms of non-parametric estimation. In essence, the kernel is used as a weighting function centred at the data points and its extension (around the data points) is defined by a smoothing parameter. A recent review of probability distributions for wind speed modelling can be found in Jung and Schindler (2019).

Linear variables characterizing the wave climate, such as significant wave height H_S and peak period T_p derived from a wave spectrum, are essential for the estimation of the design wave loads on ocean and coastal structures. For instance, the operational performance of a wave energy converter depends on the wave period while its survivability on wave height. In the context of analysing these wave parameters through univariate probability models, various models have been applied, including among others the Lognormal (Athanassoulis et al., 1994; Haver, 1985), the three-parameter Weibull (Burrows and Salih, 1986; Soares and Henriques,

1996; Vanem, 2016), the Gamma (Muraleedharan et al., 2009) and the Beta distributions (Ferreira and Soares, 1999).

On the other hand, the analysis and modelling of directional (angular) variables belongs to the realm of directional statistics that is dissimilar to the traditional linear statistics. In order to accurately analyse and model directional data, directional statistical distributions have been developed, with the von Mises (vM) distribution being among the most commonly used for unimodal samples, which is equivalent to the normal distribution for linear data; see, e.g. Mardia and Jupp (2009). For multimodal samples, finite mixtures of vM distributions have been successfully applied in many studies with respect to wind and wave direction. For instance, in Soukissian (2014), a finite mixture model of vM mixture distributions was applied for modelling wind and wave direction from *in situ* measurements at three locations of deep and intermediate water depths off the coasts of the United States. The same distribution type has been proposed by Carta et al. (2008a) and Masseran et al. (2013) for the representation of wind regimes with varying prevailing wind directions regarding two case studies in the Canary Islands (Spain) and the study of wind energy potential for nine wind stations located in Peninsular Malaysia, respectively. Another popular technique for generating circular distributions is by wrapping a linear distribution around the unit circle, giving rise to many wrapped versions of traditional probability models on the real line, such as Wrapped Normal, Cauchy (Kato and Jones, 2013), Gamma (Coelho, 2011), Lognormal, Weibull (Sarma et al., 2011) and t−distributions (Pewsey et al., 2007). Alternatives to the classical parametric models, for circular data exhibiting multimodality and asymmetry, include the kernel density estimators that were firstly considered by Hall et al. (1987) for spherical data, and Bai et al. (1988) and Fisher (1989) for directional data. The Wrapped Gaussian kernel model has been applied to ocean wave directional data by Athanassoulis and Belibassakis (2002).

1.2.1   Parametric and non-parametric models for linear variables

<u>Parametric models</u>

In the context of analysing linear data, parametric models have a dominant role in the development of statistical inference. Such models are based on certain assumptions about the examined dataset and represent just an approximation of the stochastic dynamics that generated it. For instance, a specific pdf model is assumed and the corresponding unknown parameters are estimated from the available dataset regarding inference on the unknown density model of a linear (or directional) random variable (rv). The precision of the fitting depends on various factors (e.g. sample size, area of sampled data, evaluation criteria); however, a failure in the assumption leads to completely misleading conclusions.

As already mentioned, the range of mathematical models that have been examined in recent years to describe metocean characteristics is broad. Apart from the simple univariate pdfs, the mixture distributions, which are a linear combination of two or more pdfs with appropriate weighting factors, have already been applied to various scientific fields. In this section, an overview of the most representative univariate models, as regards metocean modelling and analysis, for linear characteristics is presented. As regards the basic notation of this chapter, upper case letters are used for (linear and directional) rv's (e.g. X, Θ) and lower case letters for particular realizations of rv's (e.g. x, θ).

<u>Gamma (three-parameter) distribution (GAM)</u>

The Gamma distribution is particularly useful for modelling long-tailed and positively skewed data that can be encountered among others in hydrology and reliability studies. The pdf of a rv X following a Gamma distribution is

$$f_{\text{GAM}}(x;\, c, a, b) = \frac{(x-c)^{a-1}}{\Gamma(a)b^a} e^{-\frac{x-c}{b}}, x > c;\, a, b > 0, \tag{1.1}$$

with a and b representing the shape and scale parameters, respectively. Also, c is the location parameter and $\Gamma(\ )$ is the Gamma function defined as $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}\mathrm{d}x$.

The corresponding cdf is

$$F_{\text{GAM}}(x;\, c, a, b) = \frac{1}{\Gamma(a)} \gamma\left(a, \frac{x-c}{b}\right), \tag{1.2}$$

where $\gamma\left(a, \frac{x-c}{b}\right)$ is the incomplete Gamma function defined as $\gamma(s, x) = \int_0^x t^{s-1}e^{-t}\mathrm{d}t$.

The expected value and the variance of a Gamma rv X are $\mu = c + ab$ and $\sigma^2 = ab^2$, respectively. Several works can be found for the reliable estimation of the unknown parameters for the three-parameter gamma distribution; see, e.g. Balakrishnan and Wang (2000); Bowman and Shenton (2002); Tzavelas (2009). The Gamma distribution is also a generalization of the exponential distribution for $a = 1$.

Generalized Extreme Value distribution (GEV)

The pdf of the Generalized Extreme Value distribution is

$$f_{\text{GEV}}(x;\, c, a, b) = \begin{cases} \dfrac{1}{b} e^{\left\{-\left(1+a\frac{x-c}{b}\right)^{-1/a}\right\}} \left(1 + a\dfrac{x-c}{b}\right)^{-(1+a)/a} & , \text{for } a \neq 0 \\ \dfrac{1}{b} e^{\left\{-\frac{x-c}{b} - e^{-(x-c)/b}\right\}}, & \text{for } a = 0, \end{cases} \tag{1.3}$$

where $c, a \in \mathbb{R}$ denote the location and shape parameters, respectively, and $b > 0$ the scale parameter and with the following domain of definition

$$\begin{cases} 1 + a\dfrac{x-c}{b} > 0, \text{for } a \neq 0 \\ x \in \mathbb{R}, \quad\quad\quad \text{for } a \neq 0. \end{cases}$$

According to the value of the shape parameter, three types of distributions can be derived from Eq. (1.3), namely the Gumbel distribution for $a = 0$ (type I), the Fréchet (or inverse Weibull) distribution for $a > 0$ (type II), and the Weibull distribution for $a < 0$ (type III); for more details, see also Section 3.1.

The corresponding cdf is

$$F_{\text{GEV}}(x;\, c, a, b) = \begin{cases} e^{\left\{-\left(1+a\frac{x-c}{b}\right)^{-1/a}\right\}}, \text{for } a \neq 0 \\ e^{\left\{-e^{-(x-c)/b}\right\}}, \quad\quad \text{for } a = 0. \end{cases} \tag{1.4}$$

The unknown parameters of the Generalized Extreme Value distribution can be estimated by the maximum likelihood method (Katz et al., 2002) and the probability weighted moments method (Hosking et al., 1985); for a review of the methods proposed in the relevant literature, see Soukissian and Tsalis (2015).

Johnson S$_B$ distribution (JSB)

The Johnson S$_B$ distribution was introduced by Johnson (1949) and is one of the three families of the Johnson's system of distributions. The pdf of a Johnson S$_B$ variable X is

$$f_{\text{JSB}}(x;\, c, a, d, b) = \frac{d}{\sqrt{2\pi}} \frac{a}{(x - c)(c + b - x)} e^{-\frac{1}{2}\left[a + d\ln\left(\frac{x-c}{c+b-x}\right)\right]^2},$$
$$x \in [c, c + b];\, b, d > 0;\, c, a \in \mathbb{R}, \tag{1.5}$$

c and b denote, respectively, the location and scale parameters, while a and d affect the shape of the distribution. In particular, skewness is increased while α is increasing, in absolute value, and kurtosis is increased with increasing d. Eq. (1.5) is characterized by a bounded domain of the variable and flexibility in the distribution form due to the two shape parameters, rendering it applicable in many fields like meteorology (Tang and Lin, 2013) and hydrology (Wakazuki, 2013; Cugerone and De Michele, 2015).

The corresponding equation of the cdf is

$$F_{\text{JSB}}(x;\, c, a, d, b) = \Phi\left(a + d\ln\left(\frac{x - c}{c + b - x}\right)\right), \tag{1.6}$$

with $\Phi(\cdot)$ denoting the Gaussian cdf defined as $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} \mathrm{d}t$.

The parameters of this distribution can be estimated by applying the maximum likelihood method, the method of moments, the percentile method and (linear or nonlinear) regression methods (Phien and Jivajirajah, 1984; Scolforo et al., 2003; Zhang et al., 2003).

Lognormal (three-parameter) distribution (LGN)

The three-parameter Lognormal distribution is useful for modelling positively skewed and long-tailed data. The pdf of a Lognormal variable X is

$$f_{\text{LGN}}(x;\, c, a, b) = \frac{1}{(x - c)a\sqrt{2\pi}} e^{\left\{-\frac{[ln(x-c)-b^2]}{2a^2}\right\}}, x \in (c, \infty);\, a > 0;\, b \in \mathbb{R}, \tag{1.7}$$

with a, b and c representing the shape, scale and location parameters respectively.

The corresponding cdf is

$$F_{\text{LGN}}(x;\, c, a, b) = \Phi\left(\frac{\ln(x - c) - b}{a}\right). \tag{1.8}$$

The expected value and the variance of a Lognormal rv X are $\mu = c + e^{b + a^2/2}$ and $\sigma^2 = e^{2b + a^2}(e^{a^2} - 1)$, respectively. Mathematical properties of this distribution are described in Burges et al. (1975) and Johnson et al. (1995). The estimation techniques that are frequently used for the parameter estimation of the Lognormal distribution are maximum likelihood (Stevens, 1992; Hirose, 1997; Basak et al., 2009) and method of moments (Cohen and Whitten, 1980; Hoshi et al., 1984).

Kappa distribution (KAP)

The four-parameter Kappa distribution, introduced by Hosking (1994), is a generalization of many other three-parameter distributions; for instance, for $d = 1$ and $a \neq 0$ the Generalized Pareto model is obtained, for $d = 0$ and $a \neq 0$ the Generalized Extreme Value model, for $d = -1$ and $a \neq 0$ the Generalized Logistic distribution and for $d = 1$ and $a = 0$ the Exponential distribution. Its pdf is defined as

$$f_{\text{KAP}}(x; c, a, d, b) = \frac{1}{b}\left(1 - \frac{a(x-c)}{b}\right)^{(1/a)-1}[F_{\text{KAP}}(x)]^{1-d}, b > 0, \tag{1.9}$$

with the following domain of definition

$$\begin{cases} c + b(1 - d^{-a})/a \leq x \leq c + b/a\,, & \text{if } d > 0, a > 0; \\ c + b\log d \leq x < \infty, & \text{if } d > 0, a = 0; \\ c + b(1 - d^{-a})/a \leq x < \infty\,, & \text{if } d > 0, a < 0; \\ -\infty < x \leq c + b/a\,, & \text{if } d \leq 0, a > 0; \\ -\infty < x < \infty, & \text{if } d \leq 0, a = 0; \\ c + b/a \leq x < \infty, & \text{if } d \leq 0, a < 0. \end{cases}$$

The corresponding cdf is

$$F_{\text{KAP}}(x; c, a, d, b) = \left[1 - d\left(1 - \frac{a(x-c)}{b}\right)^{1/a}\right]^{1/d}. \tag{1.10}$$

By applying the method of L-moments for the estimation of the unknown parameters, the kappa distribution has been frequently used in hydrological studies including extreme value analysis (Park and Jung, 2002; Murshed et al., 2014; Kjeldsen et al., 2017).

Wakeby distribution (WAK)

The five-parameter Wakeby distribution, initially introduced by Landwehr et al. (1979a); Landwehr et al. (1979b), is a generalization of other less complex, such as Generalized Pareto and three-parameter Exponential distributions, and it has a great variety of shapes making it particularly useful for various applications. The pdf is given by using the following relation provided by Johnson et al. (1995)

$$f_{\text{WAK}}(x; c, g, a, d, s) = \frac{1 - \left(1 - F_{\text{WAK}}(x)\right)^{s+1}}{g\left(1 - F_{\text{WAK}}(x)\right)^{a+s} + d}, a + s > 0 \text{ or } a = d = \\ s = 0; a = 0, \text{if } g = 0; s = 0, \text{if } d = 0; d \geq 0; a + s \geq 0; g + d \geq 0, \tag{1.11}$$

with the following domain of definition

$$\begin{cases} c \leq x < \infty, & \text{if } s \geq 0, d > 0; \\ c \leq x \leq c + \dfrac{g}{a} - \dfrac{d}{s}, & \text{if } s < 0 \text{ or } d = 0. \end{cases}$$

where c, g, a are the shape parameters and d, s are the location parameters.

The corresponding cdf has no explicit analytic form and is defined by means of its quantile function

$$x(F) = c + \frac{g}{a}[1 - (1-F)^a] - \frac{d}{s}[1 - (1-F)^{-s}], F = F_{\text{WAK}}(x). \qquad (1.12)$$

The main techniques for the estimation of the parameters are maximum likelihood, method of moments and probability weighted moments (Griffiths, 1989; Rao and Hamed, 2000) while Öztekin (2011) proposed least squares method.

Weibull (three-parameter) distribution (WEI)

The three-parameter Weibull distribution, a generalization of the two-parameter Weibull distribution, has an extended range of applications due to its high flexibility; for an extensive review of applications; see Murthy et al. (2004). The pdf of a three-parameter Weibull random variable X is

$$f_{\text{WEI}}(x; c, a, b) = \frac{a}{b}\left(\frac{x-c}{b}\right)^{a-1} e^{-\left(\frac{x-c}{b}\right)^a}, x \in [c, \infty); a, b > 0; c \in \mathbb{R}, \qquad (1.13)$$

with a, b and c representing shape, scale and location parameters, respectively. The corresponding cdf is

$$F_{\text{WEI}}(x; c, a, b) = 1 - e^{-\left(\frac{x-c}{b}\right)^a}. \qquad (1.14)$$

The most common method for the estimation of the Weibull parameters is the maximum likelihood estimation (Balakrishnan and Kateri, 2008) although some deficiencies have been identified (Cousineau, 2009). Alternative estimation methods are the quantile estimation (Wang and Keats, 1995), the moment estimation (Cran, 1988; Bartolucci et al., 1999; Akdağ and Guler, 2018) and kernel density estimation (Marković et al., 2009).

Now, let us assume that the pdf $f(x)$ of a rv X is expressed in the form of a linear mixture as follows

$$f(x) = \sum_{j=1}^{\kappa} \omega_j f(x; \vartheta_j),$$

where the quantities ω_j, with $\sum_{j=1}^{k} \omega_j = 1$ and $0 < \omega_j < 1$, for $j = 1, \ldots, k$, denote the weighting factors corresponding to the components of the linear mixture, ϑ_j is the set of parameters corresponding to the $j-$th pdf and $f(x; \vartheta_j)$ is the pdf of the $j-$th component of the mixture. The following mixture distributions refer to this type of models. The unknown parameters of these distributions can be estimated by maximizing the corresponding log-likelihood function under the restriction that $\sum_{j=1}^{k} \omega_j = 1$.

Gamma-Weibull mixture distribution (GW)

The pdf of a Gamma-Weibull distribution is given by

$$f_{\mathrm{GW}}(x;\ a_1, b_1, a_2, b_2, \omega)$$
$$= \omega \frac{x^{a_1-1}}{\Gamma(a_1) b_1^{a_1}} e^{-\frac{x}{b_1}} + (1-\omega) \frac{a_2}{b_2} \left(\frac{x}{b_2}\right)^{a_2-1} e^{-\left(\frac{x}{b_2}\right)^{a_2}}, \qquad (1.15)$$

with $\omega \in (0,1)$ denoting the weighting parameter. The corresponding cdf is

$$F_{\mathrm{GW}}(x; a_1, b_1, a_2, b_2, \omega) = \omega F_{\mathrm{GAM}}(x; a_1, b_1) + (1-\omega) F_{\mathrm{WEI}}(x; a_2, b_2), \qquad (1.16)$$

with $c = 0$ from Eqs. (1.2) and (1.14), and $\omega \in (0,1)$ the weighting parameter.

Truncated Normal mixture distribution (NN)

For a normal density function with location and scale parameters c, b, respectively, i.e.

$$f_{\mathrm{N}}(x;\ c, b) = \frac{1}{\sqrt{2\pi b^2}} e^{-\frac{(x-c)^2}{2b^2}}, x \in \mathbb{R}; b^2 > 0; c \in \mathbb{R}, \qquad (1.17)$$

the singly truncated (from below) normal mixture distribution (for $x \geq 0$) is defined as follows,

$$f_{\mathrm{NN}}(x;\ c_1, b_1, c_2, b_2, \omega) = \omega \frac{\varphi(x; c_1, b_1)}{I(c_1, b_1)} + (1-\omega) \frac{\varphi(x; c_2, b_2)}{I(c_2, b_2)}, \qquad (1.18)$$

with $\omega \in (0,1)$ denoting the weighting parameter and $I(c,b) = \int_0^\infty \varphi(x; c, b)\mathrm{d}x$. The corresponding cdf is given by

$$F_{\mathrm{NN}}(x;\ c_1, b_1, c_2, b_2, \omega)$$
$$= \omega \int_0^x \frac{\varphi(x; c_1, b_1)}{I(c_1, b_1)} \mathrm{d}x + (1-\omega) \int_0^x \frac{\varphi(x; c_2, b_2)}{I(c_2, b_2)} \mathrm{d}x. \qquad (1.19)$$

Weibull (two-parameter) mixture distribution (WW)

The pdf of the Weibull mixture distribution is

$$f_{\mathrm{WW}}(x;\ a_1, b_1, a_2, b_2, \omega) = \omega \frac{a_1}{b_1} \left(\frac{x}{b_1}\right)^{a_1-1} e^{-\left(\frac{x}{b_1}\right)^{a_1}}$$
$$+ (1-\omega) \frac{a_2}{b_2} \left(\frac{x}{b_2}\right)^{a_2-1} e^{-\left(\frac{x}{b_2}\right)^{a_2}}, \qquad (1.20)$$

where a_1, b_1 are, respectively, the shape and scale parameters of the first Weibull component, a_2, b_2 are the corresponding parameters of the second Weibull component and ω is the weighting parameter. The corresponding cdf is given by

$$F_{\mathrm{WW}}(x;\ a_1, b_1, a_2, b_2, \omega) = \omega F_{\mathrm{WEI}}(x; a_1, b_1) + (1-\omega) F_{\mathrm{WEI}}(x; a_2, b_2), \qquad (1.21)$$

with $c = 0$ from Eq. (1.14) and $\omega \in (0,1)$ the weighting parameter.

Weibull-Generalized Extreme Value mixture distribution (WGEV)

The Weibull-Generalized Extreme Value mixture distribution is defined as follows

$$f_{\text{WGEV}}(x;\,a_1,b_1,c,a_2,b_2,\omega) = \omega\frac{a_1}{b_1}\left(\frac{x}{b_1}\right)^{a_1-1}e^{-\left(\frac{x}{b_1}\right)^{a_1}}$$
$$+(1-\omega)\frac{1}{b_2}e^{\left\{-\left(1+a_2\frac{x-c}{b_2}\right)^{-1/a_2}\right\}}\left(1+a_2\frac{x-c}{b_2}\right)^{-(1+a_2)/a_2}, \tag{1.22}$$

where a_1, b_1 are the shape and scale parameters, respectively, of the Weibull distribution, and a_2, b_2 and c are the shape, scale and location parameters, respectively, of the Generalized Extreme Value distribution. Its cdf is

$$F_{\text{WGEV}}(x;\,a_1,b_1,c,a_2,b_2,\omega) = \omega F_{\text{WEI}}(x;a_1,b_1)$$
$$+(1-\omega)F_{\text{GEV}}(x;c,a_2,b_2), \tag{1.23}$$

with $c = 0$ from Eq. (1.14) and $\omega \in (0,1)$ the weighting parameter.

Non-parametric models

Non-parametric methods, contrarily to the parametric ones, do not rely on strong parametric assumptions but rather on fewer, or less stringent, conditions, rendering this estimation method more flexible since the pdf is constructed according to the information derived from the available data sample and not defined by a finite set of parameters. Non-parametric techniques tend to be preferred in cases where the pdf of the data is unknown or cannot be easily approximated (e.g. due to the small sample size), and can be useful when analysing data with outliers, which might be nontrivial with a parametric approach. Nevertheless, non-parametric methods are not as optimal as parametric methods in case the assumptions of the latter ones hold.

Kernel density function (kdf), as a member of a non-parametric approach, is widely applied in non-parametric statistical estimation in data analysis and other research areas. The main idea of the kernel function is to act as a local weighting by attributing at each random variable X a weight based on the distance of observations y_i, $i = 1, \dots, n$, from a random sample to each fixed point $x \in \mathbb{R}$; the local weight increases with decreasing distance. In the univariate case and given n observations y_i, $i = 1, \dots, n$, the general form of the standard kdf, originated from Rosenblatt (1956), is defined as

$$f_{\text{K}}(x;\,y,h) = \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-y_i}{h}\right), \tag{1.24}$$

where $K(\cdot)$ is the kernel (or window) function determining the shape of the weighting function, h is the bandwidth (smoothing parameter or window width) of the estimator, which is a positive parameter that represents the variance of the kernel and thereby controls the smoothness of the estimator, and n is the sample size.

Any symmetric function that satisfies the following conditions can be used as a kernel function:

    i. $K(x;y,h) \geq 0$ for all x, i.e. is a non-negative function
    ii. $\int K(x;y,h)\mathrm{d}x = 1$
    iii. $\int xK(x;y,h)\mathrm{d}x = 0$, because of symmetry
    iv. $0 < \int x^2 K(x;y,h)\mathrm{d}x < \infty$, i.e. is of second order.

The most commonly used symmetric kdfs are the following: Gaussian, Uniform, Triangular, Epanechnikov and Biweight. However, such kernels are ideal provided that the support of target density f is unbounded; otherwise f_K is biased at the boundaries, also known as 'spill-over' effect. To overcome this problem in the case of bounded (from below or/and above) data, where standard symmetric kernels tend to underestimate the density at the boundaries, various methods have been suggested; for instance, transformations techniques (Marron and Ruppert, 1994), reflection and replication methods (Schuster, 1985; Muller, 1991; Karunamuni and Zhang, 2008), boundary kernels (Gasser et al., 1985) and adaptive kernels (Botev et al., 2010); a list of relevant methods on this topic can be also found in Karunamuni and Alberts (2005) and Marchant et al. (2013). In this thesis, the approach of using asymmetric (skewed) kernel functions is followed to construct the kdfs, i.e. kernels that matches with the support of f.

The most popular asymmetric kernel estimators include the Gamma kernel estimators (Chen, 2000), the Beta kernel estimators (Brown and Chen, 1999), the Inverse and the reciprocal inverse Gaussian kernel estimators (Scaillet, 2004) and the Birnbaum–Saunders kernel estimators (Jin and Kawczak, 2003). Moreover, Silverman (1986) has proposed Lognormal (and Gamma) kernels, with the shape parameter controlling the smoothing (Igarashi, 2016). Among the most appealing properties of these kernels are the increased precision of the density estimation close to the boundary, they are boundary bias free (i.e. the bias is of the order of $O(h)$ near the boundaries and inside the support) and their adaptive smoothing by variable kernel shapes according to the location of the data points y; more details on asymmetric kernels can be found in the recent book of Hirukawa (2018).

In the case of asymmetric kernels, the following additional requirements must be fulfilled so that the estimator of $f(x)$ remains asymptotical unbiased as $n \to \infty$ and $nh \to \infty$, for all x that belong to the support of $f(x)$, say S:

   i. $\lim\limits_{h \to 0} \int_S K(x; y, h) \mathrm{d}y = 1$

   ii. $\lim\limits_{h \to 0} \int_S K(x; y, h)(y - x) \mathrm{d}y = 0$

   iii. $\lim\limits_{h \to 0} \int_S K(x; y, h)(y - x)^2 \mathrm{d}y = 0$

As mentioned in Athanassoulis and Belibassakis (2002), the positioning parameter y is the solution of

$$\frac{\partial K(x)}{\partial x} = 0, \tag{1.25}$$

i.e. is defined as the most probable value of the kernel function, while the bandwidth parameter h is the standard deviation of the kdf with respect to x. Let it be noted that regarding the implementation of the kdfs, the positioning parameter y coincides with the values of the data sample while the bandwidth h can be derived by applying the L_2-distance criterion.

Based on the two-parameter Gamma distribution (with $c = 0$ in Eq. (1.1)), the expressions of the shape and scale parameters, a and b, respectively, of the Gamma kernel in terms of y and h are the following

$$a = \frac{1}{2}\left[2 + \left(\frac{y}{h}\right)^2 + \sqrt{\left(2 + \left(\frac{y}{h}\right)^2\right)^2 - 4}\right] \text{ and } b = \frac{y}{a-1}. \tag{1.26}$$

In a similar way, the Lognormal kernel, defined by means of the two-parameter Lognormal distribution (with $c = 0$ in Eq. (1.7)), has the following expressions for the shape and scale parameters, respectively,

$$a = \ln(Dy) \text{ and } b = \sqrt{\ln(D)}, \tag{1.27}$$

where D is the positive root of the equation $D^4 - D^3 - \left(\frac{h}{y}\right)^2 = 0$.

For sufficiently large n, the accuracy of the kernel density estimation is more sensitive to the bandwidth compared to the kernel function. The degree of smoothness of the estimated density is determined by the bandwidth; a small bandwidth yields a tight fit with spikes at the observations while a large bandwidth provides a smooth fit. In order to find the optimal bandwidth so that the obtained kernel density can adequately represent the underlying population, it is necessary to select a measure of distance that assesses the performance of $\hat{f}$ by comparing it with the true density f. Although many methods have been recommended, the non-parametric statistical community agrees that there may not be a convergence as regards the perfect method for the selection of the optimal bandwidth. The generally accepted performance criteria are the integrated squared error (ISE) and its expected value, the mean integrated squared error (MISE), given by

$$\text{ISE} = \int \{f_K(x) - f(x)\}^2 \mathrm{d}x \text{ and MISE} = E[\int \{f_K(x) - f(x)\}^2 \mathrm{d}x], \tag{1.28}$$

respectively.

Two different classes can be roughly distinguished, which asymptotically coincide: i) the cross-validation methods that try to minimize the former measure, and ii) the plug-in methods that try to minimize the latter measure; see also the review of Heidenreich et al. (2013). In the case of asymmetric kernels, the plug-in methods are not directly applicable because they require a pilot estimate of the bandwidth; for more details, see Loader (1999) and Jin and Kawczak (2003). Nevertheless, Hirukawa and Sakudo (2014) proposed the implementation of plug-in methods for choosing the smoothing parameter using the Gamma and the Modified Gamma kernels. On the other hand, the cross-validation methods for asymmetric kernels have been applied among others by Jeon and Kim (2013) and Marchant et al. (2013).

1.2.2   Parametric and non-parametric models for circular variables

<u>Parametric models</u>

The von Mises (vM) distribution was introduced by von Mises (1918) and is also referred to as circular normal distribution due to its analogy to the Gaussian distribution for linear variables. The vM distribution belongs to the exponential family and is defined as follows:

$$f_{\mathrm{vM}}(\theta; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(\theta - \mu)}, \theta \in [0, 2\pi), \kappa \geq 0, \mu \in [0, 2\pi), \tag{1.29}$$

where $I_0(\kappa)$ is the modified Bessel function of the first kind and zero order, i.e. $I_0(\kappa) = \frac{1}{2\pi}\int_0^{2\pi} e^{\kappa \cos\theta} \mathrm{d}\theta$ (or using a power series expansion $\sum_{l=0}^{\infty} \frac{1}{l!^2}\left(\frac{\kappa}{2}\right)^{2l}$), μ is the location parameter and κ is the concentration parameter around μ. The corresponding cdf does not have closed form and is calculated by numerical integration of Eq. (1.29), i.e.

$$F_{vM}(\theta; \mu, \kappa) = \int_0^\theta f_{vM}(\omega; \mu, \kappa) d\omega$$

$$= \frac{\theta I_0(\kappa) + 2\sum_{p=1}^\infty \dfrac{I_p(\kappa)\sin p(\theta - \mu)}{p}}{2\pi I_0(\kappa)}, \theta \in [0, 2\pi). \tag{1.30}$$

For $\kappa = 0$, the circular Uniform distribution is obtained, which is invariant under rotation and reflection (Mardia and Jupp, 2009). The pdf of this circular distribution is

$$f_U(\theta) = \frac{1}{2\pi}, \theta \in [0, 2\pi). \tag{1.31}$$

For the estimation of the von Mises parameters, usually the maximum likelihood method is performed, which is rather straightforward. An alternative method is to use non-parametric techniques, e.g. based on kernel approaches, as is presented at the end of this subsection.

From the wrapping of the Cauchy distribution around the unit circle, it results the wrapped Cauchy distribution with pdf

$$f_{WC}(\theta; \mu, \rho) = \frac{1}{2\pi}\left\{1 + 2\sum_{p=1}^\infty \rho^p \cos p(\theta - \mu)\right\}, \theta \in [0, 2\pi); \kappa \geq 0, \tag{1.32}$$
$$\mu \in [0, 2\pi),$$

where μ is the location parameter and ρ controls the concentration of the model. This distribution is symmetric and unimodal and has some desirable mathematic properties as discussed in Kent and Tyler (1988). Based on the trigonometric moments in the characteristic function of θ, a simplified expression of Eq. (1.32) is obtained by

$$f_{WC}(\theta; \mu, \rho) = \frac{1 - \rho^2}{2\pi[1 + \rho^2 - 2\rho\cos(\theta - \mu)]}, \rho \in [0,1]. \tag{1.33}$$

$WC(\mu, \rho)$ tends to the Uniform distribution as $\rho \to 0$ while as $\rho \to 1$, it tends to a point distribution at μ.

An analogous wrap around $N(\mu, \sigma^2)$ gives the wrapped Normal distribution $WN(\mu, \rho)$ with pdf

$$f_{WN}(\theta; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \sum_{m=-\infty}^\infty e^{-\frac{(\theta - \mu + 2\pi m)^2}{2\sigma^2}}, \theta \in [0, 2\pi); \sigma > 0, \tag{1.34}$$
$$\mu \in [0, 2\pi),$$

with μ denoting the location parameter and $\sigma^2 = -2\log\rho \Rightarrow \rho = e^{-\frac{\sigma^2}{2}}$. Another useful representation of Eq. (1.34) is in terms of the characteristic function of the normal distribution

$$f_{WN}(\theta; \mu, \rho) = \frac{1}{2\pi}\left\{1 + 2\sum_{p=1}^\infty \rho^{p^2} \cos p(\theta - \mu)\right\}, \rho \in [0,1]. \tag{1.35}$$

When $\sigma^2 \leq 2\pi$, the pdf of $WN(\mu, \rho)$ can be approximated adequately for $m = 0$ in Eq. (1.34) or by the first three terms of the infinite series of Eq. (1.35) when $\sigma^2 \geq 2\pi$ (Pewsey et al., 2013;

Mardia and Jupp, 2009). Similar to the Wrapped Cauchy distribution, $WN(\mu, \rho)$ is unimodal and symmetric about μ and tends to the Uniform distribution as $\rho \to 0$ while as $\rho \to 1$, it tends to a point distribution at μ.

For multimodal angular variables encountered in engineering applications, a finite mixture of vM distributions is implemented. The vM mixture pdf of a random angular variable θ is defined as the weighted sum of N simple vM distributions, i.e.:

$$f_{\mathrm{mvM}}(\theta; \mu_j, \kappa_j, \omega_j) = \sum_{j=1}^{N} \frac{\omega_j}{2\pi I_0(\kappa_j)} e^{\kappa_j \cos(\theta - \mu_j)}, \theta, \mu_j \in [0, 2\pi); \kappa_j \geq 0; \quad \omega_j \in [0,1], \tag{1.36}$$

where N is the number of components, κ_j and μ_j, $j = 1,2,\dots,N$, are the individual vM distribution parameters, and ω_j are (weighting) quantities with sum equal to one. The corresponding cdf for the von Mises mixture distribution is

$$F_{\mathrm{mvM}}(\theta; \mu_j, \kappa_j, \omega_j) = \sum_{j=1}^{N} \frac{\omega_j \left\{ \theta I_0(\kappa_j) + 2 \sum_{p=1}^{\infty} \frac{I_p(\kappa_j) \sin p(\theta - \mu_j)}{p} \right\}}{2\pi I_0(\kappa_j)}. \tag{1.37}$$

The parameters of the finite mixture von Mises model are estimated using the expectation-maximization (EM) algorithm for maximum likelihood estimation; for more details, see, for example, Ch. 4 of Jammalamadaka and SenGupta (2001), Mooney et al. (2003) and Banerjee et al. (2005).

Non-parametric models

The directional variables in terms of kernel density functions can be efficiently modelled by applying the Wrapped Normal distribution with the following expression for its density

$$f_{\mathrm{K,WN}}(\theta; h) = \frac{1}{nh} \sum_{i=1}^{n} K_{\mathrm{WN}}(\theta; \psi_i, h), \tag{1.38}$$

with $K_{\mathrm{WN}} = \frac{1}{h\sqrt{2\pi}} \sum_{m=-\infty}^{\infty} e^{-\frac{(\theta - \psi - 2\pi m)^2}{2h^2}}$, which can be accurately approximated by the three central terms of the sum, i.e. for $m = -1,0,1$, for moderate values of h (Athanassoulis and Belibassakis, 2002).

An expression analogous to Eq. (1.24) for kernels dealing with circular data in the q−dimensional sphere $\mathbb{S}^q$ was introduced by Hall et al. (1987). For the univariate case, i.e. for $q = 1$, the circular kernel density estimation from a random sample $\{\psi_i\}_{i=1}^{n}$ is given by

$$f_{\mathrm{K}}(\theta; h^*) = \frac{c_0(h^*)}{n} \sum_{i=1}^{n} M(h^* \cos(\theta - \psi_i)), \theta, \psi \in [0, 2\pi), \tag{1.39}$$

where $M(\theta; \psi, h^*)$ is the circular kernel, $c_0(h^*)$ is a constant such that f_{K} is a density and h^* is the concentration parameter with a behaviour similar to the inverse of the smoothing parameter

h. Small values of h^* lead to over-smoothed circular densities while large ones provide under-smoothed estimators (Oliveira et al., 2012).

Few studies have been dedicated to the appropriate selection of the smoothing parameter h^* in circular kernel density estimation. For instance, Hall et al. (1987) suggested the use of cross-validation bandwidths while Taylor (2008) proposed a rule of thumb for the selection of bandwidth assuming that the underlying population follows a vM distribution; however, the performance of the latter bandwidth may be rather unreliable if the involved data set exhibits characteristics such as multimodality and skewness. Oliveira et al. (2012) introduced a new plug-in rule procedure for bandwidth selection, following the simple idea proposed by Taylor (2008), that is based on the use of mixtures of vM distributions allowing thus more flexibility in the underlying model.

1.3   Bivariate case

Bivariate probability distributions seem to be a more realistic and complete approach in ocean and coastal engineering (e.g. for a detailed long-term analysis) since winds, waves and currents are generally non-independent variables. Numerous studies have been dedicated to the joint description of two linear rv's such as significant wave height and wind speed since wind and wave loads are crucial when assessing environmental risks for a marine structure (Nerzic and Prevosto, 2000; Zhai et al., 2017), and wave height and period as the sea state at a particular location depends primarily on these two wave parameters (Haver, 1987; Ochi, 1992; Muraleedharan et al., 2015; Vanem, 2016). On the other hand, the joint description of linear and circular variables is gaining ground since the knowledge of directionality and the corresponding bivariate stochastic structure (including one linear and one directional variable) has proved to be essential, among others, for optimizing the layout of an offshore wind farm (Feng and Shen, 2015), and for the design of coastal and offshore structures (in terms of safety, stability, strength, etc.) (Jia, 2011; Wei et al., 2017).

In this thesis, the construction of the bivariate distribution functions of the examined metocean parameters is accomplished through three different families of distributions in the parametric case, i.e. two parametric copulas, the Farlie-Gumbel-Morgenstern and the Plackett families of distributions, and the Johnson-Wehrly model, and the multiplicative kdf in the non-parametric case. A common feature to all bivariate models (parametric and non-parametric) is that their density functions rely on the corresponding univariate marginal distributions, which are known beforehand (coming from the marginal data). Moreover, all parametric bivariate models, apart from the marginal distributions functions, rely also on an additional parameter that quantifies the correlation/dependence of the variables.

In marine energy related applications, the Farlie-Gumbel-Morgenstern family has been implemented by Erdem and Shi (2011) for the estimation of the bivariate distribution of wind speed and direction, and Qu and Shi (2010) applied the same family for the joint description of wind speed and air density while Carta and Mentado (2007) examined the same variables by applying the Plackett family. Other studies where the Plackett family has been implemented in applications related to wind energy assessment are those of Carta and Velázquez (2011) for the joint description of wind speed at a candidate and reference site in the context of developing a new Measure-Correlate-Predict methodology, and Bai et al. (2016) for the joint description of wind power and wind speed while the same model has been implemented for marine applications dealing with wave data such as Athanassoulis et al. (1994), Lucas and Soares (2015) and Vanem (2016). The Johnson-Wehrly model has been implemented in Carta et al. (2008b), Qin et al. (2010), Erdem and Shi (2011), Soukissian (2014), Basile et al. (2015), Soukissian and Karathanasi (2017) and Zhang et al. (2018) for the description of wind speed and wind direction. In Soukissian (2014), the same model was also applied for the joint

description of significant wave height and wave direction. In the recent work of Han et al. (2018), the performance of the Johnson-Wehrly model was compared with the multiplicative kdf for the joint modelling of wind speed and direction based on four evaluation metrics from four wind observations sites in China. Asymmetric distribution models based on copulas were also applied by Fazeres-Ferradosa et al. (2019) to obtain the joint cumulative distribution function of significant wave height and up-crossing mean wave period.

1.3.1 Parametric models for linear-circular variables

The common characteristic of these families is that they take explicitly into consideration the marginal distribution of the corresponding variables along with their dependence structure. Well-known families of bivariate distributions are those of Fréchet (Fréchet, 1951), Johnson and Wehrly (Johnson and Wehrly, 1978), Mardia (Mardia, 1970b), Farlie-Gumbel-Morgenstern (Morgenstern, 1956; Farlie, 1960) and Plackett (Plackett, 1965) that are both particular expressions of copulas. The concept of copulas was first introduced by Sklar (Sklar, 1959) and is essentially based on the construction of multivariate distribution models characterized by the corresponding (given) marginal distributions of the involved random variables and a copula function indicating their dependence structure. This interesting characteristic of copulas (i.e. coupling the dependence of random variables with their marginal behaviours) is ideal for the construction of families of bivariate distributions (Fisher, 1997). In this section, the Johnson-Wehrly, the Farlie-Gumbel-Morgenstern and the Plackett families of distributions are described.

Johnson-Wehrly model (JW)

The joint pdf $f_{JW}(x, \theta)$ is expressed

$$f_{JW}(x, \theta) = 2\pi f_\Psi(\psi) f_X(x) f_\Theta(\theta), x \in \mathbb{R}; \theta \in [0, 2\pi), \tag{1.40}$$

where $\psi = 2\pi[F_X(x) - F_\Theta(\theta)]$, $\psi \in [0, 2\pi)$ and $f_\Psi(\psi)$ is the pdf of the rv defined by the previous equation, which represents the dependence structure between the rv's X and Θ.

Following Carta et al. (2008b) and Soukissian (2014), $f_\Psi(\psi)$ is a rather smooth function that can be described through a vM mixture pdf comprising of two components. A distinct feature of JW model is the fact that it is constructed directly for the joint description of linear and angular variables, whereas Farlie-Gumbel-Morgenstern and Plackett families are general-purpose bivariate distributions. The elegant and closed form expression of JW model, provided by Eq. (1.40), along with the fact that any marginal distribution can be considered, renders the corresponding bivariate distribution very appropriate candidate for the description of the wind/wave climate in an area. A relative drawback of this model is that the corresponding bivariate cdf can only be numerically estimated, since there is no analytic form.

Farlie-Gumbel-Morgenstern model (FGM)

The Farlie-Gumbel-Morgenstern model belongs to the family of FGM copulas that was first introduced by Morgenstern (1956) and extended by Farlie (1960); see also Ch. 44 of Kotz et al. (2000) for a more detailed introduction to multivariate distributions, including FGM distributions. A thorough presentation of the corresponding theoretical background can be also found in Athanassoulis et al. (1994).

The bivariate pdf of the FGM model is given by

$$f_{\text{FGM}}(x,\theta) = f_X(x)f_\Theta(\theta)\{1 + r_{\text{FGM}}[2F_X(x) - 1][2F_\Theta(\theta) - 1]\}, x \in \mathbb{R};$$
$$\theta \in [0, 2\pi),$$
(1.41)

where r_{FGM} is the statistical association parameter between the rv's X and Θ. Let it be noted that the bivariate FGM distribution is valid when there is a relatively weak dependence between the examined variables; in turn, the permissible range of the linear-circular correlation coefficient $r_{X\Theta}$ between X and Θ is $-1/3 \le r_{X\Theta} \le 1/3$ (Long and Krzysztofowicz, 1992; Guven and Kotz, 2008).

The corresponding bivariate cdf is provided through the following relation

$$F_{\text{FGM}}(x,\theta) = F_X(x)F_\Theta(\theta)\{1 + 3r_{\text{FGM}}[1 - F_X(x)][1 - F_\Theta(\theta)]\}.$$
(1.42)

The sample version of r_{FGM} is

$$r_{\text{FGM}}^2 = \frac{r_{Xc}^2 + r_{Xs}^2 - 2r_{Xc}r_{Xs}r_{cs}}{1 - r_{cs}^2},$$
(1.43)

where

$$\begin{cases} r_{Xc} = \rho[(x_1, \cos\theta_1), (x_2, \cos\theta_2), .., (x_n, \cos\theta_n)] \\ r_{Xs} = \rho[(x_1, \sin\theta_1), (x_2, \sin\theta_2), .., (x_n, \sin\theta_n)] \\ r_{cs} = \rho[(\cos\theta_1, \sin\theta_1), (\cos\theta_2, \sin\theta_2), .., (\cos\theta_n, \sin\theta_n)], \end{cases}$$
(1.44)

and ρ denotes the Pearson product-moment correlation based on the available sample. In Eq. (1.44), x_i and θ_i, $i = 1, ..., n$, denote realizations of the linear variable X and the angular variable Θ, respectively.

Plackett model (PLA)

The Plackett model is also a member of the copula families. Though more complicated than the other systems, it is selected for the present analysis since it is valid for any bivariate random variable (X, Θ) with $-1 \le r_{\text{PLA}} \le 1$. A detailed presentation of the corresponding theoretical background can be found in Athanassoulis et al. (1994).

The joint pdf $f_{\text{PLA}}(x, \theta)$ of the rv's X and Θ is given as a function of the two marginal pdfs and cdfs $f_X(x)$, $f_\Theta(\theta)$ and $F_X(x)$, $F_\Theta(\theta)$, respectively

$$f_{\text{PLA}}(x, \theta)$$
$$= \psi_P f_X(x)f_\Theta(\theta)\frac{(\psi_P - 1)[F_X(x) + F_\Theta(\theta) - 2F_X(x)F_\Theta(\theta)] + 1}{[S^2 - 4\psi_P(\psi_P - 1)F_X(x)F_\Theta(\theta)]^{3/2}},$$
$$x \in \mathbb{R}; \theta \in [0, 2\pi),$$
(1.45)

where $\psi_P > 0$, $\psi_P \ne 1$ is a 'correlation-type' parameter between marginal distributions, and S is given by

$$S = 1 + (\psi_P - 1)[F_X(x) + F_\Theta(\theta)].$$
(1.46)

The corresponding bivariate cdf is

$$F_{\text{PLA}}(x,\theta) = \frac{S - \sqrt{S^2 - 4\psi_P(\psi_P - 1)F_X(x)F_\Theta(\theta)}}{2(\psi_P - 1)}. \tag{1.47}$$

The estimation of ψ_P parameter can be made with various methods (Kotz et al., 2000). For the estimation of this parameter from the available data, a numerical approach is provided by the maximum likelihood estimator (Nelsen, 2006). An alternative solution is based on the cross-product ratio that provides the following estimate for ψ_P

$$\psi_P = \frac{p_1 p_4}{p_2 p_3}, \tag{1.48}$$

where $p_1 = Pr[X \leq x, \Theta \leq \theta]$, $p_2 = Pr[X \leq x, \Theta > \theta]$, $p_3 = Pr[X > x, \Theta \leq \theta]$, and $p_4 = Pr[X > x, \Theta > \theta]$ are the observed frequencies of the corresponding cells in the $(x, \theta)-$plane.

An attractive property of the Plackett model refers to the estimation of ψ_P from the observed frequencies of the four quadrants, determined by the lines that are parallel to the axes and pass through the sample medians of the two rv's X and Θ (Mardia, 1970a; Nelsen, 2006). Eq. (1.48) can be written as

$$\tilde{\psi}_P = \frac{p_1' p_4'}{p_2' p_3'}, \tag{1.49}$$

where p_1', p_2', p_3', p_4' are defined like the observed frequencies in Eq. (1.48), but x and θ are substituted by the two sample medians $\text{med}(x)$ and $\text{med}(\theta)$, respectively. Moreover, Mardia (1970a) proved that this estimator minimizes the variance of $\tilde{\psi}_P$.

1.3.2 Non-parametric models for linear-circular variables

The extension of the univariate kernel density in the bivariate case is achieved by means of the multiplicative (or product) kernels. This specific family of multiplicative kdfs is based on the product of the univariate kdfs, provided as follows:

$$f_{\text{K}}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d} K_{jh_j}(x_j - X_{ij}), \tag{1.50}$$

where $K_{jh_j}(\cdot)$ is the kernel density in the $j-$th component of the $d-$variate df with bandwidth h_j; see also Sec. 2.9 of Härdle (1991). In the bivariate case and based on the examined linear and directional variables in this thesis, the above equation is simplified to

$$f_{\text{K}}(x,\theta) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} K_X(X; X_i, h) K_{\text{WN}}(\Theta; \Theta_j, h^*), \tag{1.51}$$

where K_X can be the Gamma or the Lognormal kernel.

Chapter 2 Regression analysis & calibration of linear and directional variables

2.1 General

The collection of metocean information at an offshore/nearshore location can be achieved by measuring devices (e.g. oceanographic buoy), remote-sensing devices and numerical models at frequent temporal intervals and appropriate spatial scales depending on the purposes of the study. However, each data source is characterized by strengths and limitations, as described in Appendix E, due to the inherent uncertainties of each source. *In situ* measurements provide currently metocean data of the highest quality available; thus, they are regarded as 'ground truth' and are used for validation against the other two data sources in the area of interest, after appropriate temporal and/or spatial collocation (Schmidt et al., 2017; Young et al., 2017). Usually, prior to validation lies calibration.

The impetus for this analysis derives from the necessity of calibrating the less accurate data sources rigorously in order to obtain more reliable information as a prerequisite in a range of applications, such as wind and wave climatology, investigation of trends and design of marine structures. For instance, over- or under-estimation of wind and wave variables leads to changes in the estimation of the corresponding potential and the extreme values. Taking for granted that all the above data sources contain errors, calibration techniques should not rely on conventional linear regression analysis, which is the primary theoretical background for such applications, due to violation of assumptions. Thus, a more realistic and proper approach is to consider a regression model that takes into account errors in both variables in order to provide a statistical relationship between the less reliable data source (predictand) and the more accurate one (predictor). The so-called measurement error model serves towards this goal (Section 2.2) while robust regression models are addressed as they are less sensitive in the presence of outliers (Section 2.4), which may also distort the results of a regression model.

Despite the significance of the accurate determination of directional variables in ocean engineering applications, calibration techniques are rarely adopted for this type of variable. In this thesis, the method described for the correction of directional characteristics is based on the simultaneous minimization of the vertical and horizontal distances from each point to the regression line (Sections 2.5 and 2.6). The performance of each examined regression/ calibration model proposed in this thesis is assessed in real data samples of wind and wave data; see Section 4.3.

2.1.1 Background

Fundamentally, regression analysis is about understanding how the conditional distribution of a random variable Y changes for samples determined by possible values of one or more rv's X. Mathematically, the regression of Y on k rv's $X_1, \dots, X_k$ is written in the form

$$E[Y|X_1, \dots, X_k] = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k$$
$$= \beta_0 + \sum_{i=1}^{k} \beta_i X_i,$$

where Y, the variable of interest, is called response, predictand or dependent variable and X_i, $i = 1, \dots, k$, the explanatory variable, is called covariate, regressor, predictor or independent.

The use of terms 'dependent' and 'independent' is avoided so as to prevent any confusion with dependence/independence of rv's in the probability sense.

Typically, the aim of a regression model is to describe the statistical relation, if exists, between variables or predict the response variable for values of the covariates. Obviously, the regression model should not be viewed as the "true" model (i.e. a model that completely explains the variation in the response variable); it is most realistic to accept that a regression model verifies our theory about which variables strongly influence the response variable. Given a data sample of independent observations, the parameters that characterize the relation of these variables are usually estimated by the least-squares or maximum likelihood method.

For the sake of convenience and clarity, first some notation is introduced that is necessary before going through the various regression models that is based on the distinction of the covariates to fixed constants and random values. Observed, also called manifest or indicator, variables are denoted by uppercase Roman letters, say X or Y, while unobserved, also called true, variables are denoted by uppercase Greek letters, say Ξ or H. Analogously, each realization of an observed (unobserved) variable is denoted by lowercase Roman (Greek) letter, say x_i or y_i (ξ_i, η_i), $i = 1,2, \dots, n$. Error terms, often called random or stochastic components, are denoted by the lowercase Greek letters ε and δ. The unknown regression parameters (or coefficients) are denoted by the lowercase Greek letter β_k, $k = 0,1,2, \dots, r$, depending on the number of covariates that are considered in the model.

In the standard (population) linear regression, it is assumed that the covariates, either fixed or random, are measured without error and the corresponding model takes the form

$$Y = \beta_0 + \beta_1 \Xi + \varepsilon, \tag{2.1}$$

where ε is the error term, often written as $\varepsilon = Y - E[Y|\Xi]$, which essentially contains not only the random components of the response variable but also accounts for the effects of the covariates that are not included in the regression model (Berry, 1993). When a sample of observations $(\xi_1, y_1), \dots, (\xi_n, y_n)$ is available from the population, the most frequently used method to estimate the unknown parameters β_0 (intercept) and β_1 (slope) is least-squares (LS). When certain assumptions from the regression theory hold (see Section 2.3), then the Gauss Markov Theorem ensures that the ordinary least-squares (OLS) estimators provide the best[1] linear unbiased estimates, known briefly as BLUE, for the regression coefficients. However, for linear regression models with errors in the covariates, it is known that the least squares method yields biased[2] and inconsistent estimates for the involved parameters leading to erroneous conclusions; see also p. 3-4 of Fuller (1987). On the other hand, when only the response variable is observed with error, then the estimator is unbiased.

2.2 Measurement error models

As the title of this chapter implies, in this chapter reference is made to statistical models that take into account variables containing errors of any origin, a situation that is valid almost in every discipline. In the context of this thesis, measurement is the realization of a set of operations in the field made under (unspecified) external conditions in order to quantify the value of a physical parameter by means of the appropriate equipment and material. By the term "measurement error", two generic types of errors are usually included: i) systematic (or non-random), and ii) random (or stochastic) errors. The former error is mainly attributed to an inherent inaccuracy of the system (e.g. imperfect calibration of the measuring device) yielding

---

[1] "Best" among the class of linear unbiased estimators.
[2] The bias depends on the magnitude of the measurement error and the correlation between the covariate.

shifted measurements from the true value by a constant amount, and in turn biased OLS estimators, while the latter is unpredictable from one measurement to another. Measurement error models (MEMs) are often encountered in the literature as "errors-in-variables models". MEMs have been systematically examined by numerous researchers suggesting techniques for fitting regression lines when both variables are measured with error; see, e.g. Wald (1940), Halperin (1961), Riggs et al. (1978), Klepper and Leamer (1984) and the monograph by Fuller (1987), who covered a variety of statistical techniques for measurement error models, from simple to multivariate ones, and provided examples from various areas of application.

Now, let us consider the most common case when studying relationships between two continuous variables Ξ and H, which are related with the following linear (linearity is referred to the parameters) form

$$H = \beta_0 + \beta_1 \Xi, \tag{2.2}$$

where Ξ is the predictor variable and H is the response variable while the parameters β_0, β_1 (intercept and slope, respectively) have to be estimated. Both variables represent the "true" (unobserved) variables, which are measured with error; thus, the observed random variables are

$$X = \Xi + \delta \text{ and } Y = H + \varepsilon, \tag{2.3}$$

where δ and ε are the errors, which are uncorrelated with Ξ. Because Ξ and H are observed with error, the model of the form (2.2) and (2.3) represents the MEM.

Let us assume a sample of size n, then the unobserved variables satisfy

$$\eta_i = \beta_0 + \beta_1 \xi_i, i = 1, \dots, n \tag{2.4}$$

and the equations for the actual observed variables are

$$x_i = \xi_i + \delta_i \text{ and } y_i = \eta_i + \varepsilon_i, i = 1, \dots, n. \tag{2.5}$$

Regarding the assumptions of the random errors δ_i and ε_i, it is supposed that they have mean zero and finite variances and they are uncorrelated with each other and amongst themselves, i.e.

$$
\begin{aligned}
E[\delta_i] = E[\varepsilon_i] &= 0, && \text{for all } i, \\
Var[\delta_i] = \sigma_\delta^2, \ Var[\varepsilon_i] &= \sigma_\varepsilon^2, && \text{for all } i, \\
Cov[\delta_i, \varepsilon_j] &= 0, && \text{for all } i, j, \\
Cov[\delta_i, \delta_j] = Cov[\varepsilon_i, \varepsilon_j] &= 0, && \text{for all } i \neq j.
\end{aligned}
\tag{2.6}
$$

The existence of errors in both variables Ξ and H poses a problem different from the seemingly similar simple regression model described in Section 2.3, which is actually a special case of the MEM. Substituting Ξ, H from Eq. (2.3) to (2.2), it is obtained that

$$Y = \beta_0 + \beta_1 X + (\varepsilon - \beta_1 \delta), \tag{2.7}$$

Hence, X, which is a random variable, is correlated with the error term $(\varepsilon - \beta_1 \delta)$, with

$$Cov[X, \varepsilon - \beta_1 \delta] = Cov[\Xi + \delta, \varepsilon - \beta_1 \delta]$$
$$= Cov[\Xi, \varepsilon] + Cov[\Xi, -\beta_1 \delta] + Cov[\delta, \varepsilon] + Cov[\delta, -\beta_1 \delta]$$
$$= -\beta_1 Cov[\delta, \delta] \qquad\qquad (2.8)$$
$$= -\beta_1 Var[\delta]$$
$$= -\beta_1 \sigma_\delta^2.$$

Note that in the simple regression case, $\sigma_\delta^2 = 0$. Due to the above correlation, applying ordinary least squares estimator in a MEM, yields inconsistent estimates, which are not considered reasonable for the parameters of such a model (Cheng and Van Ness, 2010).

Based on various assumptions as regards variable Ξ, three distinct models can be formed:

a) when the ξ_i's are unknown "fixed" (meaning "not random") constants, then the model is called a functional model;
b) when the ξ_i's are random variables independent and identically distributed with $E(\xi_i) = \mu$ and $Var[\xi_i] = \sigma^2 > 0$, then the model is called a structural model, and;
c) when the ξ_i's are independent random variables with different means μ_i and common variance $\sigma^2 > 0$, then the ultrastructural model is formed. From this model, functional and structural models are derived for $\sigma^2 = 0$ and $\mu_1 = \cdots = \mu_n$, respectively. In addition, the ultrastructural model reduces to the simple linear regression one if the explanatory variable is measured without error, i.e. for $\delta_i = 0$.

In practice, it is not straightforward to determine which relation is most appropriate for the situation examined. As stated in Madansky (1959), this determination depends on the type of inference, e.g. prediction problem or testing hypothesis about the parameters. In this thesis, emphasis is given on the first two models, since they are more frequently encountered in the relevant literature with numerous practical uses.

Summing up, a MEM presumes the following three models:

i. a regression model for the association of a (unobservable) regressor variable Ξ to a response variable Y;
ii. a measurement model that relates the unobservable variable Ξ to an observable variable X and assuming an additive random error δ with mean zero, and;
iii. the generating process of the values of the true variable ξ. If these values are assumed to come from a distribution then the structural estimation is adopted, otherwise, if there is no explicit assumption for the distribution of Ξ but $\{\xi_i\}_{i=1}^n$ are rather considered as sequences of fixed but unknown values, then the functional model is used.

2.2.1 Maximum likelihood estimation for the measurement error model

Maximum likelihood (ML) is probably the most widely adopted method for parameter estimation in the MEMs. ML estimates are obtained by maximizing the likelihood function with respect to the unknown parameters. In practice, differentiating the likelihood function with respect to the parameters and setting the derivatives equal to zero and solving the resulting equations is the most common way to find the ML estimates. As regards these equations, the following possibilities are present:

a) they can have a unique solution, which indeed maximizes the likelihood function;
b) they may have more than one solutions, one of which is the global maximum that provides that ML estimate, and;

c) they may have more than one solutions, none of which are a maximum (e.g. a local maximum of the likelihood function). In this case, the likelihood function may either have no maximum, but the obtained solution can be considered as an estimator, or a maximum lies on the boundary of the parameter space. In the latter case, the ML estimate exists but the estimates are not obtained by solving these equations.

Among the first authors that used ML estimation for the MEMs was Lindley (1947), who stated that likelihood equations are consistent only if there is prior information available on the regression parameters; see also Kendall and Stuart (1973), who revised the issue of parameter estimation in a model with errors in both variables. A review of the approaches used to estimate the unknown parameters of the linear structural model can be found in Gillard (2010).

2.3 Simple linear regression for linear variables

The classical model for the simple linear regression has the following formula:

$$y_i = \beta_0 + \beta_1 \xi_i + \varepsilon_i, \qquad (2.9)$$

with the error terms ε_i being independent and identically distributed such that the mean value is zero ($E[\varepsilon_i] = 0$) and the variance constant ($Var[\varepsilon_i] = \sigma^2 > 0$). Let clarify that with regression, one can never find the "true" linear model that describes the relation of interest rather an approximation of it since the analysis is done with a sample data; as regards the true model ε denotes the variation of y from the true mean value at ξ, while in the regression, the residual values (deviation between the observed values y_i of the variable Y and the predicted ones $\hat{y}_i$ based on the OLS coefficient estimates) computed from the sample approximate the errors in the population. Hereunder, all the assumptions that need to be made in regression models are summarized.

Minimizing the sum of the squared residuals $\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$, where $\hat{\varepsilon}$ represents the vertical distance between y and $\hat{y}$, and differentiating it with respect to each of the unknown parameters, the OLS estimators of β_0 and β_1 are obtained, after setting the derivatives to zero, by

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{\xi}, \text{ and} \qquad (2.10)$$

$$\hat{\beta}_1 = \frac{s_{\xi y}}{s_{\xi \xi}}, \qquad (2.11)$$

where $\bar{y}, \bar{\xi}$ are the sample means of y_i, ξ_i, respectively, $s_{\xi y} = \sum_{i=1}^{n} (\xi_i - \bar{\xi})(y_i - \bar{y})$ and $s_{\xi \xi} = \sum_{i=1}^{n} (\xi_i - \bar{\xi})^2$.

Despite the computational simplicity and directness of constructing and applying a simple linear regression model for descriptive, inferential or prediction purposes, there are restrictive sets of assumptions that should be satisfied in order to provide valid regression estimates. Most of the assumptions concerning the standard regression model (of the population) deal with the nature of the relationship between the response and explanatory variables and the behaviour of the error term as well, and are required to obtain estimators for the coefficients of the regression

model with desirable properties, such as unbiasedness[3] and efficiency[4]. As regards unbiasedness, let clarify that it is not guaranteed that the OLS estimates for the regression parameters from a single sample will coincide with the population value but that after a large number of repeated samples from a population, the mean value of the probability distribution produced for the estimates of the parameters equals the true population value.

The first assumption is about linearity. With the term "linear" it is implied that there is a linear relation between the regression coefficients β and the covariates or alternatively that the conditional means of Y fall in a straight line. On the other hand, there is no need for the terms that involve covariates to be linear (e.g. adding the term $\cos\Xi$ is acceptable).

Regardless of the distribution of the error term, another important assumption is that the error term is uncorrelated with the covariate with zero mean error, expressed as $E[\varepsilon_i|\xi_i] = 0$, and each error term ε_i has equal variance σ^2, i.e. $Var[\varepsilon_i|\xi_i] = \sigma^2$ for all $i = 1, \dots, n$, also termed as homoscedasticity. Note that this implies that the conditional variance of y given ξ is also constant. Furthermore, the error terms for any two observations $i \neq j$ must be independent with $Cov[\varepsilon_i, \varepsilon_j] = 0$, also called lack of autocorrelation. If these assumptions are violated, then the OLS estimates of the regression coefficients still remain unbiased but no longer have the lowest variance.

Moreover, the error term is independent and identically distributed with mean value zero and constant variance. However, it is mathematically convenient to additionally assume that the errors (with respect to ξ) are normally distributed, i.e. $\varepsilon_i \sim N(0, \sigma^2)$ in order to accurately estimate confidence intervals and conduct statistical tests of significance. In turn, the observations of the response variable comes from a (conditional) normal distribution with $E[y_i] = \beta_0 + \beta_1\xi_i$ and $Var[y_i] = \sigma^2$. Furthermore, if the regression coefficients have to be estimated from a small sample, then the normality assumption is essential to justify tests of statistical significance and derive that the sampling distributions of the estimated regression coefficients are asymptotically normal distributed. On the other hand, for sufficiently large samples and due to the central limit theorem, the error term, approximated by the sum of an infinite series of independent random variables, follows a normal distribution. Nevertheless, the normality assumption is not necessary to obtain estimates of the regression coefficients that are BLUE while maximum likelihood estimator coincides with the OLS estimator of β_0 and β_1 under the normality assumption regarding the error term.

The covariate is considered to be non-random, meaning that the values ξ_i, $i = 1, \dots, n$ are assumed to be measured without error. Otherwise, biased estimates will be derived if the classical regression model is applied.

2.4 Robust regression models for linear variables

Another approach that takes into account deviations from the classical regression assumptions and attempts to temper the influence of outliers and influential observations on the estimator is robust regression. This approach aims to produce estimators that are insensitive in the presence of these points; see the standard reference works of Tukey (1960); Huber (1981); Hampel et al. (1986); Rousseeuw and Leroy (1987); Staudte and Sheather (1990); Ryan (1997); Maronna et al. (2006); Huber and Ronchetti (2009). Compared with diagnostic methods, presented in Section 2.4.2, Huber (1981) mentions that robust methods are more reliable and perform better,

---

[3] An estimator $\hat{\theta}$ f a population parameter θ is called unbiased if the mean value of an infinite number of repeated random samples is equal to the parameter being estimated, i.e. if $E[\hat{\theta}] = \theta$.

[4] An unbiased estimator is called efficient if it exhibits the minimum variance within a given set of unbiased estimators.

since they provide a middle ground between rejecting and accepting a suspicious observation without the analyst's subjective decision (to keep or remove a "suspicious" observation from the data sample). The essential difference between robust regression and diagnostic methods lies in the fact that in the latter approach the outliers are first identified and removed from the data set and then OLS method is applied on the "clean" data sample while in the former outliers are identified given that robust estimates are in hand. Moreover, if someone is not interested in detecting the outliers but only in the estimation of regression coefficients, still robust methods can provide reliable estimates since outlying data will not damage the estimation.

2.4.1 Robust vs linear regression

Simple linear regression is based on the principle of the OLS method, i.e. the minimization of the sum of squared residuals. The OLS estimators have some important properties: they are linear, unbiased, efficient (they have the minimum variance) and consistent; see also Section 2.3. However, all these properties hold *inter alia* under the assumptions that the random error terms are statistically independent, normally distributed with zero mean. In practice, these conditions are rarely satisfied or even examined and normality is usually considered just as a convenient approximation. Despite the elegant properties of the OLS method, it should be highlighted that it is clearly not robust to violations of its assumptions and especially, deviations from the normality assumption. Moreover, OLS estimates are very sensitive to outliers, even in large samples, leading to inefficient and biased results. For instance, in Zaman et al. (2001), it is noted that even a small percentage of bad or deviant observations in a very large sample can change drastically the OLS coefficients and result in systematic distortions of OLS estimates.

Robust methods have been introduced to provide relatively insensitive, consistent and high efficient estimators, when there are slight violations from the standard assumptions in the assumed statistical model, and for the rational consideration of outliers in regression analysis. In this regard, the use of robust methods is essential in various applications, since outliers are present in the available data samples (see Appendix E) while some of the main assumptions of OLS are suspicious or unrealistic, e.g. the homoscedasticity assumption.

A reasonable question that may be asked at this point is why robust regression techniques are not widely used? There is a number of potential answers to this question. For instance, there are available several types of robust regression models that need to be examined in a statistical analysis in order to select the most appropriate one while some robust methods require complex analytical methods that may be unstable. As is noted in Zaman et al. (2001), some additional reasons are the following:

- There is a rather naive trust that large sample sizes make robust techniques unnecessary.
- A certainty that the outliers can be either detected by visual inspection or by identifying unusual OLS residuals.
- There is lack of expertise as regards the interpretation of results from a robust analysis and lack of knowledge of the gains available from such analysis.

2.4.2 Unusual observations

OLS estimation can be substantially altered and lead to inaccurate results in the presence of one or multiple unusual observations in the sample. Collectively, untypical observations, the so-called outliers, exhibit inconsistency with the bulk of the data and their occurrence is a very common and delicate issue in real data analysis encountered in the context of various applications, such as data cleansing, network intrusion, severe weather prediction, geographic information systems, etc.

Taking into consideration the position of an observation in a scatter plot and the variable in which it corresponds, the conspicuous observations can be grouped in three classes; see also Ryan (1997). The outliers in the explanatory variable, also called as leverage points, can tilt the least-squares line due to the large effect on the corresponding estimator. Leverage points do not necessarily have to be outliers. The outliers in the response variable can be considered as such, because they have large standardized residuals and rather large influence on the least-squares line, since they increase the magnitude of residuals. Finally, an outlier both in the response and explanatory variables may be either a point with a large standardized residual or a point that deviates from the linear relationship set by the majority of the data or both. In this case, the relationship between the two variables must be taken into account for the detection of the outlier in question. Another significant class of atypical observations are the so-called influential points, which individually or jointly excessively influence the calculated values of various estimates (e.g. estimated regression coefficients, standard errors, estimated values); Nurunnabi et al. (2016). Chatterjee and Hadi (1986) examined thoroughly the interrelationship among outliers, leverage and influential points and highlighted that outliers and leverage points are not necessarily influential, and on the other hand, influential observations may not be high leverage points and outliers.

Since outliers may seriously affect regression analysis outputs and estimation of the relevant parameters, numerous procedures have been developed for the detection and investigation of such observations in linear regression. A straightforward way is the graphical representation (through scatterplots) of residuals and hat elements of the hat matrix, which are used as intrepretators of the amount of leverage or influence, exerted on the estimated values by the response variable (Hoaglin and Welsch, 1978), while there is an ongoing research on the simultaneous display of outliers, high leverage and influential points (Imon, 2005; Menjoge and Welsch, 2010; Nurunnabi et al., 2014). Another well-known statistical approach to measure influence is through diagnostic methods, i.e. statistics generally based on classical estimates aiming to the detection of influential points from the assumed model. This family of techniques is often implemented by the same procedure: first delete each observation one at a time and in turn, examine if there is any impact on the various calculated values. The most well-known diagnostic measures are the Cook's distance, $dfbeta_i$ and $dffits_i$. However, these techniques perform poorly in the presence of multiple outliers because of masking and swamping effects[5]. To this end, more effective, but frequently computationally expensive, diagnstics have been proposed in the relevant literature; see, for example, Barrett and Gray (1997); Wisnowski et al. (2001); Leys et al. (2018); Thennadil et al. (2018). Standard books that deal with outlier analysis are Hawkins (1980); Rousseeuw and Leroy (1987); Barnett and Lewis (1994); Aggarwal (2016).

On the other hand, removing outlying data points that are legitimately present in a data sample, as is the case of extreme values that are commonly encountered in metocean data, may lead to negative effects and wrong interpretations due to the wrong selection of model. On top of that, OLS estimation method can be highly influenced by even one outlier. When there is evidence that any of the standard regression assumptions is violated due to the presence of such aberrant observations then other statistical approaches are suggested. One of the widely used methods is robustification, which is the subject of the next sections.

Prior to the presentation of robust models, it is essential to make reference to specific statistical measures assessing robustness properties of robust regression models. Each measure of robustness describes different characteristics of the procedure, therefore they perform complementarily.

---

[5] Masking is the inefficiency of identifying a set of outliers because of the presence of another set, usually neighbouring, while swamping occurs when "clean" observations are mistaken for outliers because of the presence of another group of observations, usually distant.

2.4.3 Measures of robustness

Consistency

Consider the random variable X with probability distribution $P_\vartheta \in \mathcal{P}$ and cumulative distribution function F. ϑ can be expressed as a functional $\vartheta = T(P)$, defined on $\mathcal{P}$. Based on a random sample $X_1, \ldots, X_n$ of size n, a sensible estimator of ϑ is $T(P_n)$ (or $T(F_n)$), where P_n is the empirical probability distribution

$$P_n(A) = \frac{1}{n} \sum_{i=1}^{n} I[X_i \in A], A \in \mathcal{B}, \tag{2.12}$$

where I is the indicator function. The empirical distribution function F_n, pertaining on P_n, is

$$F_n = P_n\big[(-\infty, x]\big] = \frac{1}{n} \sum_{i=1}^{n} I[X_i \le x], x \in \mathbb{R}. \tag{2.13}$$

It is known from asymptotic statistics that as $n \to \infty$, the statistical functional $T(P_n)$ converges in probability to $T(P)$.

A desirable property of a statistical estimator is to be Fisher consistent. Consider a random sample $X_1, \ldots, X_n$, with some probability distribution P, which depends on an unknown parameter ϑ. Let $\hat{\vartheta}$ an estimator of ϑ expressed as a functional of the empirical probability distribution[6] P_n, i.e. $\hat{\vartheta} = T(P_n)$. If $T(P) = \vartheta$, $\forall \vartheta \in \Theta$, $P \in \mathcal{P}$, then $\hat{\vartheta}$ is said to be Fisher consistent (Fisher, 1922).

Breakdown point

The breakdown point ε^* of an estimator expresses the maximal amount of contamination (proportion of atypical points) an estimator can withstand before it becomes essentially useless. Breakdown point can be defined in different ways; see, e.g. Hampel (1971); Donoho and Huber (1983). In this thesis, the finite-sample breakdown point is adopted. Let a random sample $X^{(0)} = (x_1, \ldots, x_n)$ of size n, from a parametric model F depending on ϑ, and T a regression estimator such that $T(X) = \hat{\vartheta}$. Now imagine that m points from the original sample are replaced by arbitrary, and rather outlying, values, with the new sample denoted by $X^{(m)}$. The breakdown point of T for the sample $X^{(0)}$ is defined as:

$$\varepsilon_n^*\big(T, X^{(0)}\big) = \min\left\{ \frac{m}{n} : \sup_{X^{(m)}} \big\| T\big(X^{(m)}\big) - T\big(X^{(0)}\big) \big\| = \infty \right\}, \tag{2.14}$$

where $\|\cdot\|$ is the Euclidean norm and $\sup_{X^{(m)}} \big\| T\big(X^{(m)}\big) - T\big(X^{(0)}\big) \big\|$ denotes the maximum bias that result from such contamination. An estimate is said to have broken down if the maximum bias is infinite. In other words, $\varepsilon_n^*\big(T, X^{(0)}\big)$ is the smallest fraction of contaminated values in the sample that can lead to values of the estimator T far from $T\big(X^{(0)}\big)$, i.e. to unreliable estimates. For large (infinite) sample size, the breakdown point is given by $\varepsilon^* = \lim_{n \to \infty} \varepsilon_n^*$. For a more detailed description of the breakdown point, see Donoho and Huber (1983); Hampel et al. (1986); Heritier et al. (2009). Practically, the highest value of the breakdown point of an

---

[6] The empirical functional is often called (sample) statistic.

estimator one can hope is 50%, because it is not possible to discriminate "good" observations from outliers for higher values, while the lowest one is 0% with the estimates bearing no contamination. A rather surprising result is that for OLS it holds that $\varepsilon_n^* = 1/n$, which tends to zero as the sample size becomes larger. This means that even one outlying observation may have significant influence on the OLS estimates of the unknown parameters.

Influence function

Whereas the BDP is a global robustness measure in the sense that it measures the maximum amount of contamination an estimator can resist, the influence function (or curve) of an estimator, another important tool, measures local robustness, i.e. quantifies how infinitesimal perturbations (at a point x) influence an estimator in large samples.

A straightforward way to assess the influence of a single observation x on a specific sample statistic T_n (e.g. mean, median) is to calculate the difference between the corresponding values with and without x. In this respect, the standardized sensitivity curve quantifies this influence as follows:

$$SC_n(x, T) = \frac{T_n(x_1, \dots, x_{n-1}, x) - T_n(x_1, \dots, x_{n-1})}{(1/n)}, \tag{2.15}$$

The influence function (Hampel, 1974) of an estimator is an asymptotic version of its sensitivity curve (Maronna et al., 2006). By means of the influence function, the robustness of a statistic T for an infinitesimal contamination at any point x, given a sample with distribution F, is expressed as follows:

$$IF(x; T, F) = \lim_{\varepsilon \to 0^+} \frac{T(F_\varepsilon) - T(F)}{\varepsilon} = \left. \frac{\partial}{\partial \varepsilon} T(F_\varepsilon) \right|_{\varepsilon=0}, \tag{2.16}$$

where $F_\varepsilon = (1 - \varepsilon)F + \varepsilon \delta_x$, with δ_x a point-mass distribution that puts all its mass at point x and ε denoting the level of contamination. For the local robustness of T, it is required that the influence function is bounded for all x.

The gross error sensitivity of T at F is the maximum absolute value of the influence function

$$\gamma^*(T, F) = \sup_x |IF(x; T, F)|, \tag{2.17}$$

which measures the worst influence on T induced by a small perturbation of F at a point x. For a finite gross error sensitivity (that is a bounded influence function), the statistic T is called B(ias)-robust at F.

Other additional concepts that are connected with influence function are local shift sensitivity and rejection point. The former term measures the influence of shifting slightly an observation from point x to y while the latter one represents a distance measure, meaning that points lying outside this distance (centre of data) have no effect on asymptotic bias (the influence function becomes zero).

Asymptotic efficiency

In robust regression, the efficiency of an estimator is expressed as the ratio of the smallest possible variance obtained using a robust regression technique divided by the one obtained from

OLS. Recall that the OLS estimators are considered to be the most known efficient estimators, if all conditions are fulfilled, since this method possesses the minimum variance. Obviously, the ratio of the ideal estimator is equal or close to unity. In the relevant literature, the emphasis is on the asymptotic efficiency. In general, the precision of an asymptotically efficient estimator tends to the theoretical limit, as the sample size increases. For an unbiased estimator, asymptotic efficiency is the limit of its efficiency as the sample size $n \to \infty$ and depends on the population (distribution).

Summarizing the above-mentioned criteria, it is concluded that the following requirements are essential for a robust estimator

- Fisher consistency;
- non-zero breakdown point, with higher values resulting to more resistant estimators;
- bounded influence function, so that a single unusual observation will have a very limited effect on the estimation;
- quite low gross error sensitivity;
- low local shift sensitivity and finite rejection point;
- high efficiency.

2.4.4 Robust estimators

In the context of regression, some additional desirable properties for the robust estimators are regression, scale and affine equivariance. Attaining these properties it is assured that that the results of the regression analysis will not alter in case of particular transformations of the data. Based on the general case of the multiple linear regression model, let $\mathbf{y}$ denote the $n \times 1$ vector of the response variable, $\mathbf{x}_i$, $i = 1, \dots, n$, the rows of a full rank $n \times p$ matrix $\mathbf{X}$ and $\boldsymbol{\beta}$ the $p \times 1$ vector of the parameters to be estimated. An estimator T is called regression equivarient, if

$$T(\mathbf{X}, \mathbf{y} + \mathbf{vX}) = T(\mathbf{X}, \mathbf{y}) + \mathbf{v}, \tag{2.18}$$

where $\mathbf{v} \in \mathbb{R}^n$ is any vector. This condition allows the selection of any arbitrary values for the vector without any consequences in the validation of the results.

A scale equivariant estimator requires that

$$T(\mathbf{X}, c\mathbf{y}) = cT(\mathbf{X}, \mathbf{y}), \tag{2.19}$$

for any constant $c \in \mathbb{R}$. The above condition practically suggests that the measurement units of the response variable (with respect to the measurement units of the explanatory variable) do not affect the fit results.

As regards the affine equivariance, the following condition is required

$$T(\mathbf{XA}, \mathbf{y}) = \mathbf{A}^{-1}T(\mathbf{X}, \mathbf{y}), \tag{2.20}$$

for any non-singular $\mathbf{A} \in \mathbb{R}^{p \times p}$.

Let it be noted that on the face of other properties (e.g. low prediction error), sometimes equivariance is sacrificed (Maronna et al., 2006).

Going back to the univariate notation, let us assume a random sample $X_1, \dots, X_n$ with probability distribution P and distribution function F, for which we want to examine how the response variable Y is related to the covariate X by traditionally assuming a linear regression model.

Since OLS estimator is prone to outlying data, the robust methods have been proposed for estimating regression coefficients in the statistical literature, which are divided into three wide classes $L-$, $M-$ and $R-$ estimators.

<u>L – estimators</u>

Any estimator that is a linear combination of the order statistics written in the form $L_n = \sum_{i=1}^{n} c_{ni} X_{n:i}$, where c_{ni} are real constants and $X_{n:1} \leq \cdots \leq X_{n:n}$ are the ordered values of the sample, is called a (classical) L – estimator. Instead of minimizing the root mean square error, an alternative approach for the estimation of the regression parameters, proposed by Edgeworth (1887), is to minimize the sum of the absolute values of the residuals, i.e.

$$\min_{\hat{\beta}} \sum_{i=1}^{n} |\hat{\varepsilon}_i|, i = 1, \ldots, n. \tag{2.21}$$

Through the above form the least absolute deviation (LAD) estimates are achieved. LAD estimator is also referred to as the L_1 – estimator (due to the L_1 norm), while OLS is sometimes called L_2 – estimator. It is proved that the L_1 – estimator can deal with y −outliers of a sample, but remains weak against x −outliers, which have greater influence on the fitting. Because of that effect, the breakdown point this estimator will tend to zero. A more generalized method of the L_1 – estimator was proposed by Koenker and Bassett (1978) and is called quantile regression. This type of regression minimizes a sum of appropriately weighted distances between the observed values and the predicted ones through the check function $\rho_\tau(\hat{\varepsilon})$ defined by $-(1 - \tau)\hat{\varepsilon}$ if $\hat{\varepsilon} < 0$ (over-prediction) and $\tau\hat{\varepsilon}$ if $\hat{\varepsilon} > 0$ (under-prediction). For the special case where $\tau = 0.5$, the quantile regression coincides with L_1 – estimator.

A more robust alternative was first introduced by Hampel (1975) and aims at minimizing the median of the squared residuals, which is formulated as

$$\min_{\hat{\beta}} med \hat{\varepsilon}_i^2, i = 1, \ldots, n. \tag{2.22}$$

yielding the least median of squares (LMS) estimator. LMS is resistant to both $x-$ and $y-$outliers, possesses the highest possible breakdown point and is equivariant as regards linear transformations on the covariates. However, LMS performance in terms of asymptotic efficiency is rather poor.

Another known L – estimator that has higher breakdown point is the least trimmed squares (LTS) regression estimator developed by Rousseeuw (1984). This estimator is defined by

$$\min_{\hat{\beta}} \sum_{i=1}^{h} (\hat{\varepsilon}^2)_{i:n}, i = 1, \ldots, n. \tag{2.23}$$

where $(\hat{\varepsilon}^2)_{1:n} \leq \ldots \leq (\hat{\varepsilon}^2)_{n:n}$ are the squared ordered residuals (first squared and then ordered) and h, known as coverage, is the number of the remaining observations after the trimming, which should satisfy that $n/2 < h \leq n$. There is a trade-off as regards h as for small values of h a higher breakdown point is attained whereas high values of h lead to higher efficiency (on condition that the sample is not too contaminated). The maximum breakdown point (50%) is attained when $h = n/2 + [(p + 1)/2]$, with p denoting the number of explanatory variables in the regression model, which coincide with the half sample in the simple regression model; see Rousseeuw and Leroy (1987), p. 132, Theorem 6. However, in realistic applications, the proportion of outliers in the sample is much smaller – usually 10–25%. Obviously, when $h =$

n, which corresponds to the OLS estimator, the breakdown point is zero. Moreover, LTS satisfies the properties of regression, scale and affine equivariance. Although LTS has a relatively low asymptotic efficiency, but better than LMS, as stated by Croux et al. (1994), it still plays a role in the estimation of the parameters of other more robust methods.

M – estimators

In Huber (1964), the use of another approach of robust regression was introduced, the M – estimators, which are a trade-off between the efficiency of OLS and the resistance of L_1 – estimators. This class of estimators can also be regarded as a generalization of MLE. An M – estimator is defined by minimizing the following function of the residuals:

$$\min_{\hat{\boldsymbol{\beta}}} \sum_{i=1}^{n} \rho(\hat{\varepsilon}_i), i = 1, \dots, n, \tag{2.24}$$

where $\rho(\cdot)$ is called objective function, and should be continuous, non-negative, symmetric function with a unique minimum at zero. The M – estimator is simply a general robust case that results in the OLS estimator by appropriately defining function $\rho(\cdot)$.

Differentiating Eq. (2.24) with respect to the regression coefficients and setting to zero, it is obtained that

$$\sum_{i=1}^{n} \psi(\hat{\varepsilon}_i)\, \mathbf{x_i} = \mathbf{0}, i = 1, \dots, n, \tag{2.25}$$

where $\psi(\cdot)$ is the derivative of $\rho(\cdot)$, called the score (or influence) function. Eqs. (2.24) and (2.25) are not necessarily equivalent; for instance, Eq. (2.25) may have more solutions than Eq. (2.24). If a bounded, monotone $\psi-$function is chosen, then the breakdown point of this estimator is approximately 50%, leading also to bounded $\rho-$functions with unique solutions for the corresponding M – estimator. M – estimation can be robust if a $\psi-$function with rejection of remote outliers is chosen while a more reliable solution can be achieved when a redescending $\psi-$function is used, which discard completely but gradually (avoiding abrupt jumps) the effect of distant outlying observations. The essence behind the redescending M – estimators is to give maximum weight for the residuals lying around the neighbourhood of zero and the more they depart from the centre, the weight gets smaller. Among the most commonly used objective functions for M – estimators are: i) the Huber estimator; ii) the Andrew estimator; iii) the Welsch estimator, and; iv) the biweight or Tukey's bisquare estimator. See also Table 2-1.

Because M – estimator is not scale equivariant, the residuals have to be standardized by means of a preliminary (specified) estimate of scale S as follows

$$\sum_{i=1}^{n} \psi(\hat{\varepsilon}_i/S)\, \mathbf{x_i} = \mathbf{0}, i = 1, \dots, n, \tag{2.26}$$

A popular robust estimator for the scale factor S is the normalized median absolute deviation (MAD), defined by

$$S = \frac{\text{MAD}}{0.6745}, \tag{2.27}$$

where $\text{MAD} = med[|\hat{\varepsilon}_i - med(\hat{\varepsilon}_i)|]$, $i = 1, \ldots, n$. Through the constant $1/0.6745 = 1/\Phi^{-1}(0.75)$ a consistent estimator for σ is achieved for observations randomly sampled from a normal distribution. Alternatively, the regression coefficients have to be estimated simultaneously with scale. Since this estimator is based on the median, it is highly resistant to outlying observations, with $\varepsilon_n^* = 50\%$ and a bounded influence function. Generally, M – estimators are statistically better than OLS with regard to resistance and robustness to y −outliers. In some cases, their performance is poor compared to the latter, since they do not consider leverage points. In order to deal with this drawback, some authors proposed to enhance the definition of M – estimator by an appropriate weight function. To this end, Mallows (1975) proposed the generalized M – estimator by replacing Eq. (2.26) with

$$\sum_{i=1}^{n} w_i \psi(\hat{\varepsilon}_i / S) \, \mathbf{x_i} = \mathbf{0}, i = 1, \ldots, n, \tag{2.28}$$

where $w_i = W(\hat{\varepsilon}_i / s)$. In practice, the M – estimates cannot be computed directly from the data, because the weights depend upon the residuals, which in turn, depend upon the estimates. As a result, they are computed using an iteratively reweighted least squares (IRLS) algorithm.

Table 2-1. Objective and weight functions for the most popular M – estimators. r_H, r_A, r_W and r_B are tuning constants used appropriately to provide high efficiency in the normal case

Estimator	Objective function	Weight function
Least-squares	$\rho_{\text{LS}}(\hat{\varepsilon}) = \hat{\varepsilon}^2$	$w_{\text{LS}}(\hat{\varepsilon}) = 1$
Huber	$\rho_{\text{H}}(\hat{\varepsilon}) = \begin{cases} \hat{\varepsilon}^2/2, & |\hat{\varepsilon}| \leq r_{\text{H}} \\ r_{\text{H}}|\hat{\varepsilon}| - r_{\text{H}}^2/2, & |\hat{\varepsilon}| > r_{\text{H}} \end{cases}$	$w_{\text{H}}(\hat{\varepsilon}) = \begin{cases} 1, & |\hat{\varepsilon}| \leq r_{\text{H}} \\ r_{\text{H}}/|\hat{\varepsilon}|, & |\hat{\varepsilon}| > r_{\text{H}} \end{cases}$
Andrew	$\rho_{\text{A}}(\hat{\varepsilon}) = \begin{cases} r_{\text{A}}(1 - \cos(\hat{\varepsilon}/r_{\text{A}})), & |\hat{\varepsilon}| \leq r_{\text{A}}\pi \\ 2r_{\text{A}}, & |\hat{\varepsilon}| > r_{\text{A}}\pi \end{cases}$	$w_{\text{A}}(\hat{\varepsilon}) = \begin{cases} \sin(\hat{\varepsilon}/r_{\text{A}})/(\hat{\varepsilon}/r_{\text{A}}), & |\hat{\varepsilon}| \leq r_{\text{A}}\pi \\ 0, & |\hat{\varepsilon}| > r_{\text{A}}\pi \end{cases}$
Welsch	$\rho_{\text{W}}(\hat{\varepsilon}) = \frac{r_{\text{W}}^2}{2}\left[1 - \exp\left(-\left(\frac{\hat{\varepsilon}}{r_{\text{W}}}\right)^2\right)\right], |\hat{\varepsilon}| \leq \infty$	$w_{\text{W}}(\hat{\varepsilon}) = \exp\left(-\left(\frac{\hat{\varepsilon}}{r_{\text{W}}}\right)^2\right), |\hat{\varepsilon}| \leq \infty$
Biweight	$\rho_{\text{B}}(\hat{\varepsilon})$ $= \begin{cases} r_{\text{B}}^2/6\{1 - [1 - (\hat{\varepsilon}/r_{\text{B}})^2]^3\}, & |\hat{\varepsilon}| \leq r_{\text{B}} \\ r_{\text{B}}^2/6, & |\hat{\varepsilon}| > r_{\text{B}} \end{cases}$	$w_{\text{B}}(\hat{\varepsilon}) = \begin{cases} [1 - (|\hat{\varepsilon}|/r_{\text{B}})^2]^2, & |\hat{\varepsilon}| \leq r_{\text{B}} \\ 0, & |\hat{\varepsilon}| > r_{\text{B}} \end{cases}$

S – estimators

S – estimated, introduced by Rousseeuw and Yohai (1984), were developed to improve the efficiency of both LMS and LTS, based on the estimates of scale. Specifically, their objective functions are replaced by a more efficient scale estimator that is applied to the residuals $\hat{\varepsilon}_i$ in order to minimize their dispersion, and their mathematical expression is

$$\min_{\hat{\boldsymbol{\beta}}} S\big(\hat{\varepsilon}_1(\boldsymbol{\beta}), \ldots, \hat{\varepsilon}_n(\boldsymbol{\beta})\big), \tag{2.29}$$

The dispersion $S\big(\hat{\varepsilon}_1(\boldsymbol{\beta}), \ldots, \hat{\varepsilon}_n(\boldsymbol{\beta})\big)$ should satisfy the following constraint:

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{\hat{\varepsilon}_i}{S}\right) = \kappa, i = 1, \ldots, n, \tag{2.30}$$

where κ is a constant, taken often equal to $E_\Phi[\rho(\hat{\varepsilon})]$ to assure consistency of S at the standard normal distribution function Φ. S – estimators are regression and scale equivariant, and can achieve high breakdown point with the appropriate selection of ρ −function. However, they cannot combine simultaneously high relative efficiency (approximately 30%); see Hossjer

(1992). Another drawback refers to the optimization procedure implemented to obtain S – estimators, which is rather delicate. The reason lies in the fact that many local minima may emerge for the bounded, but not convex, ρ −function resulting in random resampling algorithms for the solution. This implies that the repetition of this procedure (for the same data sample) may provide different estimates. As we shall see immediately afterwards, S – estimators are usually used to provide an initial estimate in more complex robust regression methods. In order to overcome the low efficiency of S – estimators, Croux et al. (1994) proposed the generalised S – estimators, which are based on the minimization of the generalized M – estimator of residual scale.

_MM_ – estimators

Another class of robust estimators, introduced by Yohai (1987) for the linear regression setting, which tries to combine the high efficiency of M – estimators with the high resistance to outliers of S – estimators (i.e. high breakdown point), is the so-called _MM_ – estimates. There is a three-step procedure to estimate the regression coefficients $\widehat{\boldsymbol{\beta}}$: i) at the first stage, an initial estimate $\widehat{\boldsymbol{\beta}}^{(0)}$ is calculated so that it possesses a high breakdown point, such as LTS or S – estimates (with Huber or bisquare functions), but not necessarily high efficiency, ii) at the second stage, a robust M – estimate of scale S of the residuals is computed based on the initial estimate, and iii) in the final stage, $\widehat{\boldsymbol{\beta}}$ is defined as any solution of

$$\sum_{i=1}^{n} \psi\big[\hat{\varepsilon}_i\big(\widehat{\boldsymbol{\beta}}\big)/S\big]\, \mathbf{x_i} = \mathbf{0}, i = 1, \dots, n, \tag{2.31}$$

which satisfies

$$\sum_{i=1}^{n} \rho\big(\hat{\varepsilon}_i\big(\widehat{\boldsymbol{\beta}}\big)/S\big) \leq \sum_{i=1}^{n} \rho\big(\hat{\varepsilon}_i\big(\widehat{\boldsymbol{\beta}}^{(0)}\big)/S\big), i = 1, \dots, n. \tag{2.32}$$

It is obvious that at the final stage an M – estimation is carried out with only one extra condition. Thus, the IRLS can be applied to compute the potential solution of Eq. (2.31) by keeping fixed the measure of the scale estimate S in each iteration. Moreover, Yohai (1987) proved that the final estimator will obtain the highest breakdown point (i.e. 50%), if an estimator with equal breakdown point is used in the first stage. The objective functions of stages 1, 2 and 3 can vary, since the two first stages are responsible for breakdown point and the third one for asymptotic efficiency. Generally, _MM_ – estimator performs well except in the presence of high leverage points (Simpson and Montgomery, 1998).

_R_ – estimators

Robust regression estimators based on ranks of the residuals are called R – estimators, and were proposed by Hodges and Lehmann (1963) and extended by Jaeckel (1972) and others. Let R_i denote the rank of the i −th residual $\hat{\varepsilon}_i$, and $a(\cdot)$ a monotone score function such that it satisfies $\sum_{i=1}^{n} a_n(i) = 0$, then the minimization of the following sum provides the R – estimates

$$\min_{\widehat{\boldsymbol{\beta}}} \sum_{i=1}^{n} a_n(R_i)\hat{\varepsilon}_i, i = 1, \dots, n. \tag{2.33}$$

In contrast with the M – estimators, this class of estimators are scale equivariant. On the other hand, most of the R – estimates have a breakdown point $\varepsilon_n^* = 1/n \to 0$, when $n \to \infty$. R – estimators are not used in this thesis, but are described in brief for completeness.

2.5 Circular regression models

Circular regression regards the prediction of the circular response variable conditional on the explanatory one and can be found in numerous applications including engineering and environmental sciences. Regression models that handle circular data are generally categorized in three groups depending on the nature of the variables involved. In this respect, i) if both the response and the explanatory (or covariate) reside on the unit circle, we refer to a circular-circular regression model, ii) if the response variable takes values on the real line and the explanatory on the unit circle, then the regression model is called linear-circular, and iii) circular-linear regression for the vice versa case.

An early study of regression models including circular variable(s) was made by Gould (1969), who introduced a maximum likelihood solution for estimating the parameters of a (multivariate) circular-linear regression problem; however, those parameters were not unique (Lund, 1999; SenGupta and Ugwuowo, 2006). Some years later, in the works by Mardia (1972), Johnson and Wehrly (1978) and Fisher and Lee (1992) improvements of Gould's model were presented while Laycock (1975) described several regression models including circular variates. Lund (1999) defined a regression model for the prediction of a circular variable by a circular predictor and a set of linear covariates showing that least circular distance and maximum likelihood estimates coincide if circular response follows a von Mises distribution. SenGupta and Ugwuowo (2006) studied a multivariate regression model with a linear response variable, a circular explanatory one, expressed as a trigonometric polynomial, and a set of linear covariates. This model was applied to solar and wind energy data.

Circular-circular regression models have been proposed by many authors. For instance, Jammalamadaka and Sarma (1993) introduced a circular regression model for two circular rv's, where the circular response variable is expressed through sine and cosine functions that are regressed on functions expressed in terms of the Fourier series expansions of the circular explanatory variable. The estimation of parameters of the suggested model is based on least squares. Rivest (1997) provided a circular regression method for predicting direction using a rotation of the decentred predictor with application to earthquake datasets. In the work by Downs and Mardia (2002), the proposed regression model between circular variables was based on a tangent link function and is equivalent to the Möbius circle transformation on the complex plane. The latter mapping was also adopted by Kato et al. (2008) for the introduction of a new circular-circular regression model, under the assumption that errors follow a Wrapped Cauchy distribution instead of the von Mises distribution, which was applied to marine biology and wind direction data. The two latter regression models were extended by Kato and Jones (2010) and Hussin et al. (2004), who proposed a linear association between the two circular variables by constraining for practical reasons the real-valued parameter of the explanatory variable to take values close to unity. Polsen and Taylor (2015) after presenting a review on circular-circular regression models, introduced a method for the detection of influential observations. SenGupta and Kim (2016) proposed the least circular distance estimation method in order to analyse the circular variables of a new circular-circular regression model in the context of determining the relationship of circular genomes.

For the rest of this section let θ_i and ϕ_i be the observed values of the circular explanatory variable Θ and the circular response variable Φ, respectively. Since the function atan2 returns values within the interval $(-\pi, \pi]$, it is more convenient to work with directions in this interval.

The mapping to $[0, 2\pi)$ can be easily made by using the modulo operation of each direction with 2π.

2.5.1 Circular-circular regression model

In order to obtain rational results after performing a circular regression, it is necessary to include a link function between the circular variables as the concept of scaling is non-existent; in this respect, various approaches have been developed for circular regression (Rueda et al., 2016). In this thesis, the circular-circular regression model presented by SenGupta et al. (2013) and SenGupta and Kim (2016), which is a generalization of Downs and Mardia's model, is adopted. First, let us consider the following mapping

$$\tan\left(\frac{\Theta - \mu_\Theta}{2}\right) = \beta_0 + \beta_1 \tan\left(\frac{\Phi - \mu_\Phi}{2}\right), \qquad (2.34)$$

where μ_Θ and μ_Φ are the location parameters of the circular variables Θ and Φ, respectively, β_0 is a real number denoting the rotation from μ_Θ, and β_1 is the slope regression parameter in the closed interval $[-1, 1]$. See also Figure 2-1, for a graphical representation of this mapping.

Since the tangent function has double solutions in $(-\pi, \pi]$ and the range of arctangent function is limited to $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$, a unique solution of Eq. (2.34) can be obtained if we turn to half angles, i.e.

$$\Theta = \mu_\Theta + 2 \tan^{-1}\left\{\beta_0 + \beta_1 \tan\left(\frac{\Phi - \mu_\Phi}{2}\right)\right\}. \qquad (2.35)$$

Now let us suppose that θ_i's are the sample values from the response circular variable Θ, which are subject to error. Each of these values are observed for each fixed value ϕ_i of the explanatory circular variable Φ. Based on the above mapping, the circular-circular regression model is defined as follows:
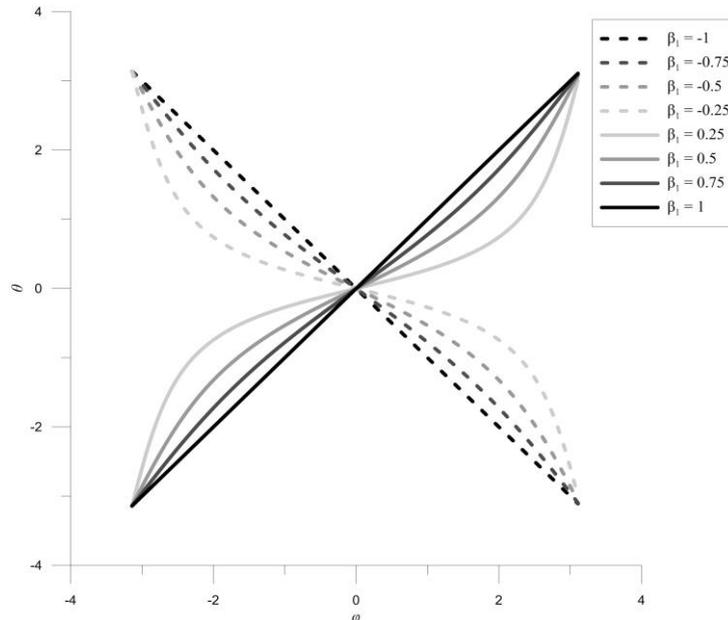
Figure 2-1. Graphical representation of the tangent mapping for selected of β_1. μ_Θ, μ_Φ and β_0 are set to zero.

$$\theta_i = \mu_\theta + 2\tan^{-1}\left\{\beta_0 + \beta_1\tan\left(\frac{\phi_i - \mu_\phi}{2}\right)\right\} + \varepsilon_{\theta,i}, i = 1, \dots, n, \tag{2.36}$$

where $\varepsilon_{\theta,i}$ is a random error that follows the von Mises distribution with zero mean direction and a constant concentration parameter.

Replacing μ_ϕ by $\bar{\phi}$, the estimation of the unknown parameters of the model in Eq. (2.36) (β_0, β_1 and μ_θ) is based on the minimization of the circular distance, a non-negative measure, between two angles ψ, ω expressed as follows:

$$d(\psi, \omega) = 1 - \cos(\psi - \omega), d \in [0, 2]. \tag{2.37}$$

For $|\psi - \omega| = \pi(\mathrm{mod}2\pi)$, then $d = 2$, while $d = 0$ implies that $\psi = \omega(\mathrm{mod}2\pi)$. The minimization of the sum of squared distances applied in the linear regression is replaced by the minimization of sums of the above circular distance given by the following form

$$\begin{aligned}
\min D(\beta_0, \beta_1, \mu_\theta) &= \min \sum_{i=1}^{n}\left[1 - \cos(\theta_i - \hat{\theta}_i)\right] \\
&= \min \sum_{i=1}^{n}\left[1 - \cos\left(\theta_i - \mu_\theta - 2\tan^{-1}\left\{\beta_0 + \beta_1\tan\left(\frac{\phi_i - \bar{\phi}}{2}\right)\right\}\right)\right],
\end{aligned} \tag{2.38}$$

where $D(\cdot)$ is the sum of the circular distances. Taking the first order partial derivatives of $D(\cdot)$ with respect to the parameters β_0, β_1 and μ_θ, it is obtained

$$\begin{aligned}
\frac{\partial D(\beta_0, \beta_1, \mu_\theta)}{\partial \beta_0} &= \sum_{i=1}^{n}\left[\frac{2\sin\left(\theta_i - \mu_\theta - 2\tan^{-1}\left\{\beta_0 + \beta_1\tan\left(\frac{\phi_i - \bar{\phi}}{2}\right)\right\}\right)}{1 + \left(\beta_0 + \beta_1\tan\left(\frac{\phi_i - \bar{\phi}}{2}\right)\right)^2}\right] \\
&= 0;
\end{aligned} \tag{2.39}$$

$$\begin{aligned}
&\frac{\partial D(\beta_0, \beta_1, \mu_\theta)}{\partial \beta_1} \\
&= \sum_{i=1}^{n}\left[\frac{2\tan\left(\frac{\phi_i - \bar{\phi}}{2}\right)\sin\left(\theta_i - \mu_\theta - 2\tan^{-1}\left\{\beta_0 + \beta_1\tan\left(\frac{\phi_i - \bar{\phi}}{2}\right)\right\}\right)}{1 + \left(\beta_0 + \beta_1\tan\left(\frac{\phi_i - \bar{\phi}}{2}\right)\right)^2}\right] = 0;
\end{aligned} \tag{2.40}$$

$$\frac{\partial D(\beta_0, \beta_1, \mu_\theta)}{\partial \mu_\theta} = \sum_{i=1}^{n} -\sin\left(\theta_i - \mu_\theta - 2\tan^{-1}\left\{\beta_0 + \beta_1\tan\left(\frac{\phi_i - \bar{\phi}}{2}\right)\right\}\right) = 0. \tag{2.41}$$

2.5.2 Linear-circular and circular-linear regression models

When modelling the relation between a linear response variable Y and a circular explanatory one Θ (case of linear-circular regression), a simple model, proposed by Mardia (1976), can be written of the form

$$y_i = \beta_0 + \beta_1\cos\theta_i + \beta_2\sin\theta_i + \varepsilon_{y,i}, i = 1, \dots, n, \tag{2.42}$$

where y_i denotes the linear observations, θ_i is the circular observations and $\varepsilon_{y,i}$ are the random errors assumed to be independent and identically distributed following a normal distribution with mean 0 and (constant) variance σ^2. The regression coefficients $\beta_0, \beta_1, \beta_2$ can be estimated by applying the least squares or the maximum likelihood methods.

A similar model has been proposed by Kim and Sengupta (2015) with the following form

$$y_i = \beta_0 + \beta_1 \cos(\theta_i - \mu_\theta) + \varepsilon_{y,i}, i = 1, \dots, n, \tag{2.43}$$

where μ_θ denoted the mean direction. For $\beta_1 > 0$ and θ moving away from μ_θ, y decreases, while it increases as θ is moving towards μ_θ. For $\beta_1 < 0$, the opposite is valid. Let it be noted that there is 2-to-1 mapping from θ to y (or from y to θ); e.g. y takes the same value when $\theta = \mu + \pi$ and $\theta = \mu - \pi$.

Likewise, in case a circular response variable Θ is to be modelled as a function of a linear explanatory variable Y, the following model can be applied

$$\cos(\theta_i - \mu_\Theta) = \beta_0 + \beta_1 y_i + \varepsilon_{\theta,i}, i = 1, \dots, n, \tag{2.44}$$

assuming that $\varepsilon_{\theta,i}$ follows the von Mises distribution with zero mean 0 and concentration parameter κ.

2.6   Calibration models

The classical statistical calibration is used when someone wants to predict an estimate of a new $\hat{x}_0$ given a new observed y_0. Assuming the simple case of model (2.9), the calibration model is written as (Eisenhart, 1939)

$$\hat{x}_{0,C} = -\frac{\hat{\beta}_0}{\hat{\beta}_1} + \frac{1}{\hat{\beta}_1} y_0, \tag{2.45}$$

with $\hat{\beta}_0$ and $\hat{\beta}_1$ denoting the estimated parameters from the regression method.

The assumptions that hold for the above model is that the regressor X is measured without error while the random errors of Y are normally distributed about the true values, have constant variance σ^2 and are independent of X. As was demonstrated by Williams (1969), the reciprocal of the slope, assuming Cauchy-like behaviour, has infinite variance and thus, infinite mean squared error[7].

An alternative approach is to regress X on Y, a procedure known as inverse regression, and has the following form

$$\hat{x}_i = \beta_0' + \beta_1' y_i + \omega_i, \tag{2.46}$$

where y_i and ω_i are independent and also $\omega_i \sim \mathcal{N}(0, \sigma_\omega^2)$. Hence, the estimate $\hat{x}_0$ is provided by

$$\hat{x}_{0,I} = \hat{\beta}_0' + \hat{\beta}_1' y_0. \tag{2.47}$$

---

[7] The variance (and bias) of the slope and the mean squared error are important statistical properties in linear regression since other properties depend on them (e.g. variance of intercept).

However, the inverse regression assumes that Y is measured with negligible error, which is usually unrealistic in real-world applications. The major differentiation between the two methods is that the former one minimizes the sum of the vertical distances between measurements and the fitted line while in the latter one the sum of the horizontal distances is minimized.

Krutchkoff (1967) compared inverse and reverse regression by means of Monte Carlo simulations and showed that the reverse regression is more efficient regarding prediction based on the mean squared error; however, his approach has come under some criticism (Berkson, 1969; Halperin, 1970; Osborne, 1991) and supports as well (Centner et al., 1998; Srivastava, 1995; Tellinghuisen, 2000). These two calibration methods can lead to different estimates of the regressor variable X and raised many controversies; yet no definitive solution has been obtained due to the complexity of the problem (Kannan et al., 2007).

Although a wide variety of studies have been focused on the classical calibration and inverse regression for linear variables, there are very limited works on the calibration problem for circular variables despite their apparent value in offshore and nearshore applications. A reference work for this subject is provided by SenGupta et al. (2013).

Let clarify that the calibration procedure in this thesis is implemented in order to correct the response variable in terms of the regressor one rather than predict the regressor variable after observing one or more values of the response variable. This concept has been adopted in a series of publications (Soukissian et al., 2014; Soukissian and Papadopoulos, 2015b; Karathanasi et al., 2016); particular results of some of them are briefly presented in Part II.

2.6.1 Calibration of linear variables

Suppose that for an unknown $\varXi_f$, Y_f is obtained, according to Eq. (2.3) by

$$Y_f = H_f + \varepsilon_f, \varepsilon_f \sim \mathcal{N}(0, \sigma_\varepsilon^2)$$
$$H_f = \beta_0 + \beta_1 \varXi_f.$$

Based on Eq. (2.1), $\varXi_f$ can be estimated by

$$\hat{\varXi}_{f,C} = -\frac{\hat{\beta}_0}{\hat{\beta}_1} + \frac{1}{\hat{\beta}_1} Y_f, \tag{2.48}$$

while in light of Eq. (2.46), the corresponding estimate is given by

$$\hat{\varXi}_{f,I} = \hat{\beta}_0' + \hat{\beta}_1' Y_f. \tag{2.49}$$

2.6.2 Calibration of directional variables

In order to apply the classical circular regression, firstly the estimation of the unknown parameters of Eq. (2.36) is necessary by minimizing the objective function $Q(\cdot)$ based on the circular distances between the initial and predicted values of the response variable $\varPhi$ (see also Eq. (2.37)):

$$\min Q(\beta_0, \beta_1, \mu_\theta, \mu_\phi) = \min \sum_{i=1}^{n} \left[1 - \cos(\theta_i - \hat{\theta}_i)\right]$$

$$= \min \sum_{i=1}^{n} \left[1 - \cos\left(\theta_i - \mu_\theta - 2\tan^{-1}\left\{\beta_0 + \beta_1 \tan\left(\frac{\phi_i - \mu_\phi}{2}\right)\right\}\right)\right]. \tag{2.50}$$

The estimates $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\mu}_\theta$, $\hat{\mu}_\phi$ are obtained by setting the first order equations for each parameter equal to zero. The calibration equation for predicting the unknown value of the regressor Φ is estimated by

$$\hat{\phi} = \hat{\mu}_\phi + 2\tan^{-1}\left\{\frac{\tan\left(\frac{\theta - \hat{\mu}_\theta}{2}\right) - \hat{\beta}_0}{\hat{\beta}_1}\right\}. \tag{2.51}$$

On the other hand, in order to implement inverse circular calibration, Eq. (2.36) is solved with respect to ϕ and the obtained relation is the following:

$$\phi = \mu_\phi + 2\tan^{-1}\left\{\frac{\beta_0}{\beta_1} - \frac{1}{\beta_1}\tan\left(\frac{\theta - \mu_\theta}{2}\right)\right\}$$

$$= \mu_\phi + 2\tan^{-1}\left\{\beta_0' - \beta_1'\tan\left(\frac{\theta - \mu_\theta}{2}\right)\right\}. \tag{2.52}$$

where $\beta_0' = \frac{\beta_0}{\beta_1}$ and $\beta_1' = \frac{1}{\beta_1}$.

The objective function that is minimized has the form

$$\min Q(\beta_0', \beta_1', \mu_\theta, \mu_\phi) = \min \sum_{i=1}^{n} \left[1 - \cos(\phi_i - \hat{\phi}_i)\right]$$

$$= \min \sum_{i=1}^{n} \sum_{i=1}^{n} \left[1 - \cos\left(\phi_i - \mu_\phi - 2\tan^{-1}\left\{\beta_0' + \beta_1'\tan\left(\frac{\theta_i - \mu_\theta}{2}\right)\right\}\right)\right]. \tag{2.53}$$

As regards the prediction of the unknown values ϕ based on the estimation of β_0', β_1', μ_θ, μ_ϕ parameters, the obtained equation is

$$\hat{\phi} = \hat{\mu}_\phi + 2\tan^{-1}\left\{\hat{\beta}_0' + \hat{\beta}_0'\tan\left(\frac{\theta - \hat{\mu}_\theta}{2}\right)\right\}. \tag{2.54}$$

When the circular calibration is based on the orthogonal distance, i.e. on the simultaneous minimization of both vertical and horizontal distances used in Eqs. (2.50) and (2.53), respectively, the corresponding objective function is

$$Q(\beta_0, \beta_1, \mu_\theta, \mu_\phi) = \sum_{i=1}^{n} \left[1 - \cos(\theta_i - \hat{\theta}_i)\right] + \sum_{i=1}^{n} \left[1 - \cos(\phi_i - \hat{\phi}_i)\right]$$

$$= \sum_{i=1}^{n} \left[1 - \cos\left(\theta_i - \mu_\theta - 2\tan^{-1}\left\{\beta_0 + \beta_1 \tan\left(\frac{\phi_i - \mu_\phi}{2}\right)\right\}\right)\right] \tag{2.55}$$

$$+ \sum_{i=1}^{n} \left[1 - \cos\left(\phi_i - \mu_\phi - 2\tan^{-1}\left\{\frac{\tan\left(\frac{\theta_i - \mu_\theta}{2}\right) - \beta_0}{\beta_1}\right\}\right)\right].$$

and is minimized for the estimation of the unknown parameters. Then, according to the obtained estimates $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\mu}_\theta$, $\hat{\mu}_\phi$, the unknown values of Φ can be predicted by applying Eq. (2.51) for new values of Θ, since these two methods rely on the same regression equation (i.e. Eq. (2.36)).

Chapter 3 Directionality in extreme value analysis of linear variables

3.1 General

In order to ensure that offshore and coastal facilities or structures are reliable both in terms of structure and economic viability, it is of paramount importance to estimate accurately the behavior of extreme values of the involved environmental variables during the structural and risk assessment. Typical metocean parameters that are analysed in the context of extreme value analysis through statistical approaches are wind speed, significant wave height and wave period, which are mainly obtained by long-term (of the order of 30 years or more) hindcast data bases and measurements with constant statistical properties in time. However, there are some factors that affect wind and wave regimes and violate the assumption of stationarity since spatial, temporal and/or directional variations may take place in the long-term study of a phenomenon. For instance, the statistical characteristics of wind speed present seasonal variations and consequently, affect the distribution parameters. The dependence of direction on the variability of a specific parameter is also evident; a typical example is the generation of higher waves at particular directional sectors compared to others at an offshore location.

In extreme value theory, there are two central approaches that are broadly used: the block maxima and peaks over threshold approaches. In the former case equal-sized non-overlapping bins (blocks) are generated to extract maximum observations while in the latter case observations above a certain threshold, appropriately selected, are extracted; however, there is not yet an established and robust methodology for the threshold selection; see, e.g. the recent review of Bücher and Zhou (2018) for an edifying discussion on these two methods. Both methods usually represent the metocean conditions in terms of amplitude and frequency assuming a unidirectional behavior for practical reasons.

Directionality is the objective of this chapter, where directional variables are incorporated in the peaks-over-threshold (POT) approach in order to investigate the dependence of extreme values of wind and wave characteristics on a directional covariate. Although the beginnings of extreme value theory date back to the 1920s with various works investigating the limiting distribution of the largest order statistic (e.g. from R. von Mises, L.H.C. Tippett, R.A. Fisher), the first attempts to incorporate directionality in extreme value models was made in the early 1980s (see, e.g. Graham (1981) and Moriarty and Templeton (1983)). In this connection, only recently directionality has been adopted as a covariate in the formulation of metocean design criteria since the dynamic behaviour and performance of marine energy devices is affected by directionality characteristics (see, e.g. Philippe et al. (2013)).

In the subsequent sections of this chapter the impacts of wind and wave directionality for specific types of wave energy devices and foundations of wind turbines at sea are discussed in order to emphasize the benefit from considering directionality features during the design of such structures. Next, a short overview of the basic concepts from the classical extreme value theory is presented along with some well-known methods for threshold selection and declustering. Then, the directional extreme model is analytically described and some modifications are introduced as regards the estimation of extreme parameters.

3.2 Response of marine energy devices to directionality

In the context of ocean energy technology, the knowledge about the extreme behaviour of wind and wave features including directional dependence is crucial; see, e.g. Larsén et al. (2015). For instance, most of the support structures for offshore wind turbines, either fixed or floating, are non-axisymmetric (apart from monopile foundations) leading to different operational response and capacity as regards loading intensity from metocean characteristics and fatigue performance. The fact that the overall cost of energy can be lowered through the continuous development of design of such structures and improved risk assessment techniques render directionality an integral part of design optimization, and reliability and safety maximization. On top of that, current regulations and standards from well-established organizations related to engineering design principles for structures at sea, such as the American Petroleum Institute (API) and the Det Norske Veritas (DNV), recommend as well the consideration of directionality to ensure proper structural safety.

The effects of directionality on some well-known wave energy converters are presented and different types of floating structures for offshore wind turbines, with low and high degree of influence from directional features, are also addressed.

3.2.1 Wave energy converters

Typically, offshore waves propagate towards a wave energy device, deployed offshore, from a range of directions thus, this device has to cope with this variability. The capture of energy from offshore waves can be achieved if the wave energy device has either a sufficient compliant mooring system that allow it to be aligned with the orientation of the device to the mean direction (for efficient power conversion) or a symmetrical frame. On the other hand, as the waves are travelling from deep to shallow waters, they are refracted while approaching the shore and they end up travelling at right angles to the shoreline regardless of the original direction of propagation. Hence, wave energy devices can be placed on the shore since the wave direction can be easily determined in advance due to this natural phenomenon.

Based on the effects of directionality in the performance and efficiency of wave energy converters (WECs) deployed offshore or nearshore, indicative examples from existing advanced technologies with different working principles, horizontal sizes and orientation are presented dealing with WECs whose performance is either highly affected by the incoming wave front or its influence is considered negligible.

WECs highly influenced by wave direction

Attenuators and terminators are the most common types of WECs whose performance is highly influenced by their orientation with respect to the prevailing direction of a given sea state. Specifically, attenuators are elongated floating devices that are oriented parallel to the wave direction, with a horizontal extent comparable to the wavelength, which lie in a predefined place thanks to moorings on the seabed. The incoming wave that passes along the device generates movements within the device that in turn exerts force on a turbine that produces energy. The most well-known attenuator is the Pelamis (http://www.emec.org.uk/about-us/wave-clients/pelamis-wave-power) which is the first commercially viable device that generated energy from the waves and provided electricity to the grid via cable. It was developed and manufactured in Scotland by an Edinburgh-based company and its first major demonstration project with three full-scale devices, with a rated power of 750 kW each, was in Aguçadoura, Portugal. The Pelamis consists of a set of semi-submerged cylinders that are linked by hinged joints (right representation of Figure 3-1). The motion of the joints (in the heave and sway

directions) induced by waves is resisted by hydraulic rams, which pump high-pressure oil through turbines driving electrical generators for power generation.
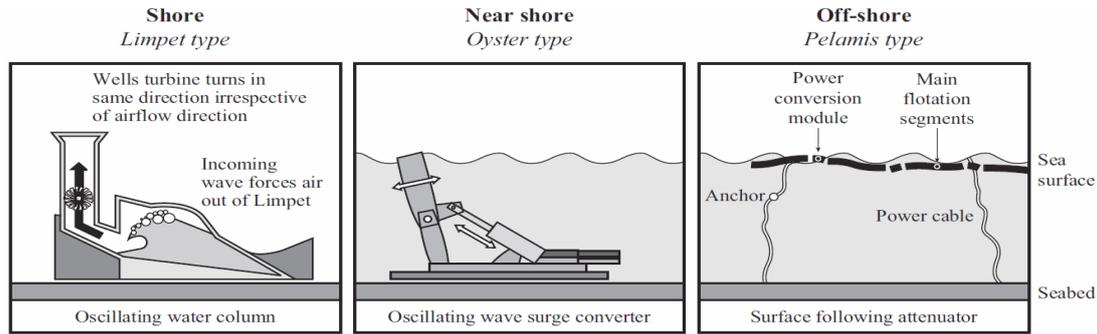
Figure 3-1. Types of WECs (Source: Tavner (2017)).

In opposition to the attenuators, terminators are oriented perpendicular to the predominant wave direction. Two typical forms of terminators are the oscillating water column (OWC) devices and the oscillating wave surge converters (OWSCs); see also left and middle representations of Figure 3-1. OWC devices consists of a partially submerged part (bottom-fixed, shore-mounted or floating) forming an air chamber. As the free surface of the water is risen inside the chamber due to the incident wave, the volume of air enclosed by this chamber is compressed. Then, the compressed air escapes through an aperture above the water column, which is connected to a bidirectional air turbine mounted on top of the structure for electricity generation. As the water inside the chamber drops, the air pressure is decreased making the air go back through the turbine. LIMPET was the first commercial-scale OWC, of 500 kW capacity, built in 2000 in the Scottish island of Islay. In contrast to OWC devices, OWSCs are driven by the horizontal particle motion of the wave, which is amplified in the near-shore environment. Oyster is an example of a flap-type OWSC. It is a near-shore device, fixed to the seabed at around 10–16m depth, and consists of a hinged mechanical flap. This flap is moved as the waves pass over the device and drives two hydraulic pistons to deliver pressurised water to an onshore turbine for the generation of electricity. The first full-scale demonstrator Oyster began producing power in 2009 when it was launched at the European Marine Energy Centre in Orkney, Scotland, where wave energy was converted to electricity through a 315kW electrical generator driven by a hydro-electric turbine.

WECs not influenced by wave direction

WECs that are able to capture energy from waves from any direction without having a principal direction like attenuators or terminators belong in the category of point absorbers. Typically, point absorbers are cylindrical in shape, with diameter smaller than a typical wavelength, and constrained to one or more degrees of freedom (usually the heave motion) while most their designs refer to a body symmetric about the vertical axis. Such devices have often relatively simple structure compared to the other wave energy converters and can extract energy from waves coming in any direction by oscillating with the movements of water for those that float near/at the surface or for submerged devices, they move up and down due to the variations in subsea pressure induced by the motion of waves. These movements can generate energy by their transfer against some sort of resistance that can take various forms.

Two representative point absorbers that have reached an advanced stage of technological development is the AquaBuOY and the Archimedes Wave Swing (AWS) with the main alteration the non-fixed and fixed bottom end of the structure to the sea bed, respectively; see also Table 1 of Bozzi et al. (2018). AquaBuOY consists of a floating buoy that is connected underneath with a large cylinder. In the center and inside this cylinder, a piston is housed and is connected with both ends of the buoy through a hose pump. This hose pump is stretched and

compressed accordingly to the relative motion between the buoy and the piston and in turn the flow of the pressurized water drives a Pelton turbine, generating power. The corresponding rated power is 250 kW. On the other hand, AWS (http://www.awsocean.com/archimedes-waveswing.html) consists of a completely submerged air-filled cylinder with a 'floating' upper part, which moves vertically with respect to the bottom-fixed part, the so-called basement. The changes in water pressure as wave crests and troughs passes over the device consecutively induce the movement of the floater, which is converted into electrical energy via a hydraulic system. A pilot plant that was deployed offshore the northern Portuguese coasts in 2004 had a rated maximum power of 2 MW (Cruz and Sarmento, 2007).

The performance of a WEC in converting energy from waves is expressed through the capture width indicator, defined as the ratio between the absorbed power and the incident wave power of a wave-front equal to the width of the WEC times this width. According to the type of WEC, the captured width ratio could vary with peak wave period but with wave direction as well. Note that even in the case where the wave energy device is less sensitive to wave direction in terms of power capture, it is possible that an array of devices is affected by the corresponding layout and the spatial orientation, and the shadow effects as well. Interactions within a wave farm can produce dissimilarities in power absorption; hence, some devices operate at their full capacity while others at a more reduced rate.

3.2.2   Floating wind turbines

Concerning the recent technology of offshore wind turbines mounted on floating support structures for water depths over 50 m, they can be roughly classified in three categories: i) semi-submersible support wind turbines; ii) spar-type support wind turbines, and; iii) tension leg platforms. The corresponding concepts are shown in Figure 3-2. In terms of affordability, the most promising floating support structures are the first two types; however, further improvements need to be considered for future large scale implementations. The first floating pilot wind farm of 30MW capacity is based on spar-type structures, which are located offshore Peterhead, Scotland. In addition, some full-scale prototype semi-submersible wind turbines have been tested at sea, e.g. WindFloat installed off the Portuguese coast at 60m water depth in 2011.

Essentially, the motion stability of a floating wind turbine highly depends on the wind and wave forces and moments acting on it, which vary in amplitude, direction and frequency over a typical design life of approximately 25 years and generate structural vibrations and extreme loads on various components of the structure. Since hydrodynamic characteristics differ along with wind and wave headings, it would be valuable to have a better understanding of the directionality effects on such structures to mitigate loads and improve efficiency. Considering the characteristics of a spar-type floating wind turbine, Barj et al. (2014) revealed that the aligned wind and wave conditions induce the highest extreme and fatigue loads for most structure locations while including misaligned wind and wave conditions can be useful to improve the estimation of extreme and fatigue loads. Moreover, Lyu et al. (2019) discovered that longitudinal modes (surge and pitch motion) of this floating system are mostly dependent on wind loads while transverse modes (sway and roll motion) rely mostly on the wave loads. As regards heave motion caused by buoyancy, it seems that it is independent of wind and wave directions.

Taking into account a triangular semi-submersible foundation, its asymmetric structural features make it sensitive to wind and wave direction resulting to diverse hydrodynamic loads. Specifically, Bachynski et al. (2014) studied the platform motions and tower loads in aligned and misaligned wind and wave conditions for two triangular semi-submersible platforms, among others, and showed that the former conditions caused the largest tower base fatigue

damages while the latter contributed to slightly increased motions. Karimirad and Michailides (2016) also concluded that motions, tension of mooring lines and functionality of a V-shaped semi-submersible floating turbine are not significantly affected when misaligned wind-wave conditions are present in operational conditions. Furthermore, Antonutti et al. (2016) showed by means of simulations that as regards wave direction, surge, heave and pitch are mainly affected by inclination for collinear wind and waves while sway, roll and yaw are altered by cross wave-and-wind cases. In Zhou et al. (2017) the performance of a Y-shaped semi-submersible wind turbine in different load directions was investigated through model tests and numerical simulations; the obtained results indicated that directionality affects the operation and maintenance of this type of floating wind turbine and the corresponding impacts should be predetermined thoroughly before installation in the offshore environment.

As regards the power efficiency of a floating wind turbine, the rotor adapts to the incident wind direction through a hydraulic pitch system ensuring that the wind direction is perpendicular to the disk formed by the rotor blade of the wind turbine as much as possible so that the captured wind energy is maximized. However, abrupt changes in wind direction cause stress on the rotor module of the wind turbine due to the constant disparities on load conditions. Furthermore, as in the case of WECs, the efficiency of an offshore wind farm is influenced by wake effects. Even small changes in wind direction can change the power output of the wind farm due to the increase of power losses making the power production less predictable.

3.3 Introduction to extreme value analysis: basic concepts and theoretical results

Let us consider a sequence of independent and identically distributed random variables $X_1, X_2, \ldots, X_n$, following the cumulative distribution function (cdf) F_X (or simply F). Let also $M_n = \max\{X_1, X_2, \ldots, X_n\}$, $n \in \mathbb{N}$, denote the maximum random variable of this sequence. In order to describe the probabilistic behavior of M_n, the corresponding cdf $F_{M_n}(x)$ should be evaluated. This can be done using the following relation:

$$F_{M_n}(x) = P(M_n \leq x) = \prod_{i=1}^{n} P(X_i \leq x) = [F(x)]^n, \text{ for } x \in \mathbb{R}, \ n \in \mathbb{N}. \qquad (3.1)$$

If the distribution F and the sample size n are known, the principal problem in extreme value theory can be solved. However, it is rare in practice that the underlying cdf F is known. Even
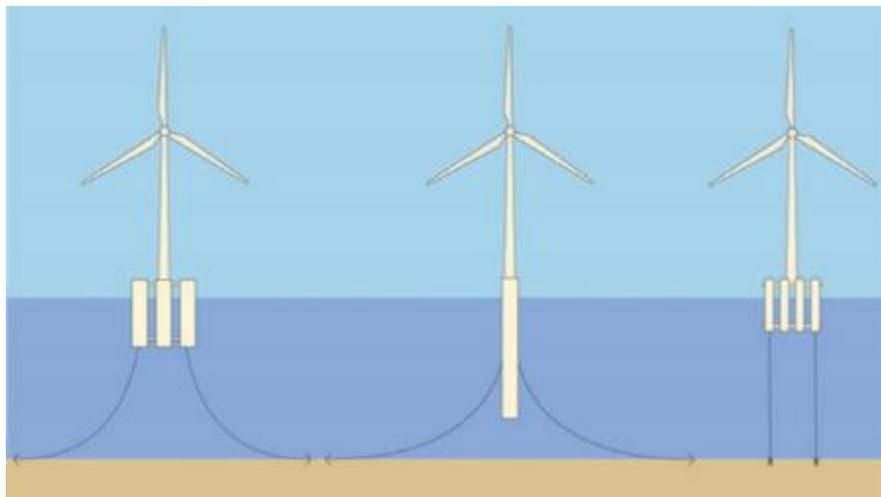
Figure 3-2. Floating wind turbine support structures: (left) semi-submersible structure; (middle) spar-type structure; and (right) tension leg platform (Wu et al., 2019).

56

if F can be estimated from a given sample and replace the theoretical one in Eq. (3.1), small discrepancies in the estimation of F can lead to significant discrepancies in F^n as n increases.

Accepting that F is unknown, it is essential to examine the asymptotic behavior of M_n near the upper endpoint of the support of F (in its right tail), where extremes occur. Let us denote $x_+ = \sup\{x \in \mathbb{R}: F(x) < 1\}$ the right endpoint of F. As $n \to \infty$, the cdf of M_n converges to a degenerate distribution on the upper endpoint x_+, which is

$$F_{M_n}(x) \underset{n\to\infty}{\to} \begin{cases} 0, & \text{for } x < x_+ \\ 1, & \text{for } x \geq x_+. \end{cases} \tag{3.2}$$

Thus, the searching of a limiting distribution $F_M(x)$ (that will depend on n) is necessary such that $F_M(x) = \lim_{n\to\infty} F_{M_n}(x)$. To avoid degeneration of $F_M(x)$, M_n is modified by means of a linear normalization expressed by $M = \sigma_n x + \mu_n$, where the quantities $\sigma_n > 0$ and $\mu_n \in \mathbb{R}$ are properly selected such that

$$F_M(x) = \lim_{n\to\infty} M_n(\sigma_n x + \mu_n) = \lim_{n\to\infty} [F(\sigma_n x + \mu_n)]^n = G(x). \tag{3.3}$$

of all $x \in \mathbb{R}$, at which G is continuous. In the context of extreme value theory (i.e. the theory for studying the stochastic structure of the rv M_n), except for finding all possible (non-degenerate) distributions G that satisfy Eq. (3.3), the distributions F have to be characterized for which sequences of $\{\sigma_n\}$ and $\{\mu_n\}$ exist such that Eq. (3.3) holds for any limit distribution. Before the formulation of the proposition that ensures the existence of the limiting distributions for M_n^*, two fundamental concepts for extreme value theory are introduced, the concepts of maximum domain of attraction and max-stability.

Let F a non-degenerate cdf (of X). F is said to belong to the *maximum domain of attraction* of cdf G, i.e. $F \in \mathcal{D}(G)$, if there exist sequences of $\{\sigma_n\}$ and $\{\mu_n\}$, such that F satisfies the relationship

$$\lim_{n\to\infty} P(M_n \leq \sigma_n x + \mu_n) = \lim_{n\to\infty} [F(\sigma_n x + \mu_n)]^n = G(x). \tag{3.4}$$

A non-degenerate cdf G is *max-stable* if there are constant μ_n and $\sigma_n > 0$, for each $n = 2,3, ...$, such that $[G(\sigma_n x + \mu_n)]^n = G(x)$. An alternative definition is derived as follows: "A non-degenerate cdf G is *max-stable* if, for each $n = 2,3, ...$, the cdf's G^n and G are of the same type." If G_1 and G_2 are two cdf's of the same type, i.e. $G_2(x) = G_1(\mu x + \sigma)$, and $F \in \mathcal{D}(G_1)$, then $F \in \mathcal{D}(G_2)$. From the above definition we see that every max-stable distribution is a limit distribution for maxima of iid rv's (Embrechts et al., 1997).

Now, the possible limiting distributions for M_n^* are provided by the theorem attributed to Fisher and Tippett (1928) and Gnedenko (1943), stating that if there exist two sequences of constants $\{\sigma_n > 0\}$ and $\{\mu_n\}$ such that $(M_n - \mu_n)/\sigma_n$ converges in distribution to G, with G a non-degenerate distribution function, then G belongs to one of the three families:

i) Gumbel family (type I): $G_G(x) = \exp\left\{-\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right\}$, $x \in \mathbb{R}, \sigma > 0$;

ii) Fréchet family (type II): $G_F(x) = \begin{cases} 0, & x \leq \mu \\ \exp\left\{-\left(\frac{x-\mu}{\sigma}\right)^{-a}\right\}, & x > \mu, a > 0; \end{cases}$

iii) Weibull family (type III): $G_W(x) = \begin{cases} \exp\left\{-\left[-\left(\frac{x-\mu}{\sigma}\right)^a\right]\right\}, & x < \mu, a > 0 \\ 1, & x \geq \mu, \end{cases}$

where $\mu \in \mathbb{R}$ and $\sigma > 0$ are the location and scale parameters, respectively, and $a > 0$ (for types II and III) is the shape parameter. The three different limit distributions are called *extreme value distributions* and the associated pdfs for types I, II and III, respectively, are given by:

$$g_G(x) = \frac{\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\exp\left\{-\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right\}}{\sigma}, x \in \mathbb{R}, \sigma > 0;$$

$$g_F(x) = \begin{cases} 0, & x \le \mu \\ \frac{a}{\sigma}\left(\frac{x-\mu}{\sigma}\right)^{-(a+1)}\exp\left\{-\left(\frac{x-\mu}{\sigma}\right)^{-a}\right\}, & x > \mu, a > 0; \end{cases}$$

$$g_W(x) = \begin{cases} \frac{\alpha}{\sigma}\left(\frac{x-\mu}{\sigma}\right)^{a-1}\exp\left\{-\left[-\left(\frac{x-\mu}{\sigma}\right)^{a}\right]\right\}, & x < \mu, a > 0 \\ 0, & x \ge \mu. \end{cases}$$

Note that every max-stable distribution is of extreme value type and conversely, each distribution of extreme value type is max-stable. In this way, it is ensured that the normalized rv M_n^* has a limiting distribution that must take only one of the above specified forms, as the sample size increases, regardless of the distribution F of the X_i, $i = 1, \dots, n$. In this sense, the Fisher-Tippett-Gnedenko theorem is an analogous result of the central limit theorem; in the latter theorem, the limit of the sums of iid rv's is described while in the former the limit of maxima is of interest.

When each rv X_n has a given cdf F, it is important to know which (if any) of the three types of limit distribution applies. von Mises (1936) established some simple sufficient conditions, such that the maxima of a distribution function F with density f converge to some specific distribution function G while various necessary and sufficient conditions, involving the "tail behaviour", are known for each type of limit; for the complete proofs, see de Haan (1976). However, in real applications the underlying assumption of iid rv's is usually not satisfied. On these grounds, Leadbetter (1974) proved that these distributions hold also for dependent rv's provided that there is long-range dependence at extreme levels

These three limit distributions can be integrated into a single parametric representation, as suggested by von Mises (1936) and Jenkinson (1955), widely known as the Generalized Extreme Value (GEV) distribution. The cdf of GEV is expressed in the following form:

$$G(x; \mu, \sigma, \xi) = \begin{cases} \exp\left[-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right], \xi \ne 0, 1 + \xi\frac{x-\mu}{\sigma} > 0 \\ \exp\left[-\exp\left(-\frac{x-\mu}{\sigma}\right)\right], \qquad \xi = 0, x \in \mathbb{R}, \end{cases} \tag{3.5}$$

with μ, σ, ξ denoting the location, scale and shape parameter, respectively.

The corresponding pdf is

$$\begin{aligned} g(x;& \mu, \sigma, \xi) \\ &= \begin{cases} \frac{1}{\sigma}\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-(1+1/\xi)}\exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, \xi \ne 0 \\ \frac{1}{\sigma}\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\exp\left[-\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right], \qquad \xi = 0. \end{cases} \end{aligned} \tag{3.6}$$

Notice that the shape parameter ξ, the so-called extreme value index, characterizes the tail behaviour of GEV distribution at its upper end, a significant feature corresponding to different behaviour of extreme events both in magnitude and frequency; i) for $\xi = 0$, the Gumbel family is obtained, which is unbounded with a tail of exponential type; ii) if $\xi > 0$, the Fréchet class is bounded from below with right heavy tail, and; iii) if $\xi < 0$, GEV is reduced to the Weibull class is bounded from above with finite upper endpoint. Summing up, the two fundamental theorems mentioned above provide a model for the description of maxima values, which in practical applications are partitioned into blocks and are called block maxima. However, the rational of block maxima method has a significant limitation; it discards important information of the extremal behaviour of a variable that usually lasts for some period of time leading to a less representative selection of extreme values. Moreover, the choice of the block size is a trade-off between bias and variance. A large block size leads to large variance in estimation due to the generation of few block maxima; on the other hand, a small block size may lead to bias due to the poor approximation of the asymptotic distribution.

An alternative parametric model for the estimation of the tail behaviour has been proposed by Pickands–Balkema–de Haan Theorem (Balkema and de Haan, 1974; Pickands, 1975) and is based on the number of excess values k over a high enough (predefined) threshold u from the initial set of observations for statistical inference. Let $X_1, X_2, \ldots, X_n$ be a sequence of iid rv's with common cdf F and let also $Y_j = X_{(j)} - u$, $j = 1, \ldots, n_u$, with $X_{(j)}$ denoting the observations that exceed the threshold u. The distribution of excesses $\left(Y_1, \ldots, Y_{n_u}\right)$ F_u can be described using the conditional probability $F_u(y) := P[Y \leq y | X > u]$. F_u can be written in terms of F as follows:

$$F_u(y) := P[Y \leq y | X > u] = \frac{P[X \leq y + u, X > u]}{P[X > u]} = \frac{F(y + u) - F(u)}{1 - F(u)}. \qquad (3.7)$$

Solving Eq. (3.7) with respect to the unconditional distribution $F(x)$ the following representation is derived:

$$F(x) = \left(1 - F(u)\right) F_u(x - u) + F(u).$$

The limit theorem by Pickands–Balkema–de Haan states that the distribution of exceedances $(X - u)$ $F_u(y)$ for large values of the threshold u can be approximated by the generalized Pareto (GP) distribution defined by:

$$H(y; \sigma_u, \xi) = 1 - \left(1 + \frac{\xi y}{\sigma_u}\right)^{-1/\xi}, \qquad (3.8)$$

for $y > 0$ and $1 + \xi y / \sigma_u > 0$, where ξ is the shape parameter and $\sigma_u > 0$ is the scale parameter, if F belongs in one of the three domains of attraction of the GEV distribution such that $\lim_{u \uparrow x_+} \sup_{0 \leq y \leq x_+ - u} |F_u(y) - H(y)| = 0$.

Hence, for large enough u, the distribution of exceedances over u is estimated by $F_u(y) \approx H(y)$. For $\xi \geq 0$ the domain of y is $[0, +\infty)$ and G follows a reparametrized version of Pareto distribution when $\xi > 0$, and an exponential distribution when $\xi = 0$, and for $\xi < 0$, it is $[0, -\sigma_u/\xi]$ and GP distribution becomes a Pareto type II distribution. Note that GEV and GP distributions share the same shape parameter ξ while the scale parameter σ_u is defined as a function of the location and shape parameters of GEV, given by $\sigma_u = \sigma + \xi(u - \mu)$. Similarly to GEV, the shape parameter of GP determines the tail behaviour: for $\xi < 0$ the distribution has an upper bound, for $\xi > 0$ the distribution has no upper limit and for $\xi = 0$ is also unbounded.

As regards the pdf of GP, it yields correspondingly:

$$h_u(y; \sigma_u, \xi) = \begin{cases} \dfrac{1}{\sigma_u}\left(1 + \dfrac{\xi y}{\sigma_u}\right)^{-(1+1/\xi)}, & \xi \neq 0 \\ \dfrac{1}{\sigma_u}\exp\left(-\dfrac{y}{\sigma_u}\right), & \xi = 0. \end{cases} \tag{3.9}$$

The examination of the rv Y leads to an alternative approach, called peaks-over-threshold (POT) method, the theoretical background of which was developed from hydrologists during the 80s; see, e.g. Todorovic (1978), Revfeim (1983). Summarizing, the GEV distribution G describes the limit distributions of the normalized maxima while the GP distribution H is the limit distribution of the excesses over a high threshold u.

3.4   Return period and return level

After assessing the adequacy of the model through goodness-of-fit tests, the return levels, associated with certain return periods, can be estimated. For a specific event, the T −year return period denotes that there is $1/T$ probability (on average) in any particular year that the T −year event will be exceeded, under the assumption of stationarity. For a given return period T of interest with an associated return level x_T that the maxima can reach, we have that:

$$\begin{aligned} F(x_T) &= P[M_n \leq x_T] \\ &= 1 - P[M_n > x_T] \\ &= 1 - 1/T. \end{aligned} \tag{3.10}$$

Hence, the return period is defined as $T = 1/[1 - F(x_T)]$.

Supposing that the GP distribution is suitable for modelling the exceedances and having estimated its unknown parameters by the ML method, from Eq. (3.7) the probability of a rv X exceeding a threshold u is

$$P[X > x] = p_u\left[1 + \hat{\xi}\left(\frac{x - u}{\hat{\sigma}_u}\right)\right]^{-1/\hat{\xi}}, \tag{3.11}$$

where $p_u = P[X > u]$, i.e. the probability of threshold exceedance. Introducing the term 'mean exceedance rate', which is the average number of observations above the threshold u per year, an estimate of p_u can be given by the empirical distribution function

$$\hat{p}_u = \frac{n_u}{n}, \tag{3.12}$$

where n_u is the number of observations exceeding the threshold u. Let it be noted that $\hat{p}_u$ is also the ML estimate of p_u, since the number of threshold exceedances follow the binomial distribution $Bin(n, p_u)$.

Now, assuming that n measurements $X_1, \ldots, X_n$ were taken during n_y observation years then it is implied that during T years there are nT/n_y observations. Thus, the x_T −return level (that is exceeded on average once in T years) is obtained by rearranging Eq. (3.11) and using Eq. (3.12).

$$x_T = \begin{cases} u + \dfrac{\hat{\sigma}_u}{\hat{\xi}}\left[\left(\dfrac{n_u T}{n_y}\right)^{\hat{\xi}} - 1\right], & \hat{\xi} \neq 0 \\[3ex] u + \hat{\sigma}_u ln\left(\dfrac{n_u T}{n_y}\right), & \hat{\xi} = 0. \end{cases} \qquad (3.13)$$

3.5   Threshold selection for excess models

The a priori selection of a suitable threshold implies the existence of an additional unknown parameter for the GP distribution, which may affect the validity of the estimates and is still an open issue with no established approach. As with the block size in the block maxima method, it is a trade-off between bias and variance. A low threshold will result in large bias and low variance leading to incorrect results for the obtained estimates since less representative extreme data are taken into account whereas a high threshold will result in small bias and large variance in the estimation of the parameters leading to unreliable results due to the smaller sample size.

A plethora of statistical techniques has been proposed for the determination of the appropriate threshold; see, e.g. the reviews of Scarrott and MacDonald (2012) and Langousis et al. (2016) for more details. According to the latter work, these methods can be roughly categorized as follows: i) graphical methods where one searches for linear behaviour of the GP parameters (or related metrics) within a range of thresholds, such as mean residual life plot and parameter stability plot; ii) goodness-of-fit-tests that detect the lowest threshold for which the GPD is suitable either by minimizing the asymptotic mean square error of the estimators or quantifying the deviations between the theoretical distribution and the empirical cdf, and; iii) non-parametric methods that determine the appropriate starting point of the extreme region of the data record. Since each method leads to different threshold choices, the sensitivity of the inferences (as regards parameter estimation) is evaluated as well. Thus, in the subsequent sections, a summary of the most widely used approaches that will be used in this thesis is presented.

Mean excess plot

Following the threshold stability property of the GP distribution (i.e. shape and modified scale parameters remain constant for higher values of the threshold) and supposing that the excesses over a threshold u^* follow this distribution, Davison and Smith (1990) suggested using the mean of the GP distribution

$$E[X - u^*|X > u^*] = \frac{\sigma_{u^*}}{1 - \xi}, \qquad (3.14)$$

for $\xi < 1$, which is called mean excess (or mean residual life) function of X. For any threshold $u > u^*$, the above expectation takes the form

$$E[X - u|X > u] = \frac{\sigma_u}{1 - \xi} = \frac{\sigma_{u^*} + \xi u}{1 - \xi}, \qquad (3.15)$$

which is linear in u with slope $\xi/(1 - \xi)$.

Given an iid sample $X_1, \dots, X_n$, an estimator of Eq. (3.15), say $\hat{e}(u)$, is the empirical mean excess function defined as:

$$\hat{e}(u) = \frac{\sum_{i=1}^{n_u}(X_i - u)I_{\{X_i > u\}}}{\sum_{i=1}^{n_u}I_{\{X_i > u\}}}, \qquad (3.16)$$

where $I_{\{X_i > u\}} = 1$ if $X > u$ and 0 otherwise, meaning that it is estimated as the ratio of the sum of the exceedances over the threshold and the total number of observations exceeding the threshold. The properties of mean excess function are described in Hall and Weller (1981). A proper threshold can be obtained by plotting $\hat{e}(u)$ as a function of the threshold u and identifying the lowest value of threshold above which $\hat{e}(u)$ increases approximately linearly. This plot has been implemented in practice by Hogg and Klugman (1984); Begueria (2005); Sanchez-Arcilla et al. (2008) among others.

Threshold stability plot

An alternative graphic technique focuses on the stability of parameter estimates for a range of threshold values u; see Section 4.3.4 of Coles (2001). If a GP model is acceptable for fitting the excesses over a threshold u^*, then for increased thresholds, e.g. $u > u^*$, the excesses should also follow a GP distribution with the same shape parameter at threshold u^* and a new scale parameter. The scale parameter σ_u is estimated by $\sigma_u = \sigma_{u^*} + \xi(u - u^*)$. The modified scale parameter can be reparametrized as $\sigma_u - \xi u$, which is constant with respect to u. Consequently, the estimates of the shape and modified scale parameters remain constant above u^*, if excesses follow the GP distribution with u^* being a valid threshold.

Estimates of the shape and the modified scale parameters are plotted against u and the appropriate threshold corresponds to the lowest threshold value for which these estimates are nearly constant. Mean excess and threshold stability plots can be applied simultaneously to obtain the optimum threshold. The main drawbacks of the above graphic approaches as a method of threshold selection is that they require expertise from the analyst for the interpretation of these diagnostics and they can be quite subjective. In addition, as a non-automated method, it is not suggested when multiple locations need to be examined in the context of extreme value analysis.

Percentiles

Among the most common rules of thumb used to derive threshold values is the percentiles. In the relevant literature, a range of percentiles have been proposed. For instance, Dumouchel (1983) suggested the upper threshold of 10%, but with inadequate theoretical justification, while Eastoe and Tawn (2012) used the 95th percentile for river flow data. Grabemann and Weisse (2008) chose to represent extreme conditions of wind speed and significant wave height by applying the 99th percentile while in Arns et al. (2013), percentiles varying between the 97.5th and the 99.7th percentile were examined in order to derive the most appropriate threshold for water level data from tide gauge records in various locations; the 99.7th percentile was identified as the most appropriate for the examined study areas.

3.6 Declustering

Regarding the extreme values of metocean parameters, it is valid that if the time step of the series is smaller than a typical duration of an extreme event (i.e. storm) then they occur in clusters, implying that there is temporal correlation between sequential values. However, in order to apply the POT method, it is essential to ensure that there is mutual independence between extreme events. The prerequisite of independence is achieved by means of

declustering, a method that takes out the dependent observations from a correlated series of extreme events so that independent threshold exceedances are extracted reasonably. This approach was implicitly applied first by Davenport (1964) and its main principle is to select the maximum value between consecutive up- and down-crossings of the mean. Several declustering techniques have been developed in the context of extreme value analysis, and the outline of this procedure is summarized below:

i. Define clusters of observations in case of consecutive exceedances based on an empirical criterion or parametric models (e.g. Markov chain models, Bartlett-Lewis process).
ii. Identify the highest value in each cluster, called declustered peaks.
iii. Assume the declustered peaks are independent and fit GPD to these peaks.

It is evident that the definition of the cluster entails some degree of subjectivity or arbitrariness, especially when empirical rules are applied, affecting in turn the results. On the other hand, in Davison and Smith (1990) it was stated that if a reasonable selection is made as regards the average number of clusters per unit time for the identification of clusters then the results seem to be insensitive to this precise value. Moreover, Ledford and Tawn (2003) introduced a diagnostic tool to evaluate the declustering of a series.

A brief overview of the most commonly used declustering methods for POT models is provided below.

Runs declustering method

Runs declustering method, described by Smith and Weissman (1994), assumes that successive threshold exceedances form a separate cluster as long as their duration does not surpass a set run length, i.e. a predefined minimum interval between two successive peaks indicating the termination of a cluster. As in the case of the threshold selection u, there is no formal procedure for the selection of run length; thus, in order to avoid improper choices of run length, which may lead to bias or high variance, the choice of the run length relies on the common sense experience and the physical background that governs the variable of interest. For instance, when studying ocean waves variables, the run length should be large enough so that the entire duration of fully developed sea states is included. In the relevant literature, a run length of 30h to 96h is chosen to ensure independence between the declustered peaks (Morton et al., 1997; Fawcett and Walshaw, 2007; Kapelonis et al., 2015; Lerma et al., 2015; Samayam et al., 2017; Santos et al., 2017).

Intervals declustering

A more sophisticated and automatic declustering scheme was developed by Ferro and Segers (2003) with the aim of determining the run length from the data. It is based on the a priori estimation of the extremal index θ_u, which represents the proportion of the times between threshold exceedances that can be considered as the times between independent clusters. A review of estimation methods for the extremal index can be found in Ferreira (2018).

The main difference with runs declustering method is that it does not involve any arbitrary choice in the process of obtaining independent clusters of exceedances and that the automation of the technique lies in the interconnection of threshold selection and declustering, meaning that a different extremal index is chosen with changes in the POT threshold. This approach has been applied by Acero et al. (2011); Cebrian and Abaurrea (2006) and Della-Marta et al. (2009) among others while it was also adopted in this thesis.

Declustering Algorithm (DeCA)

In the context of acquiring statistically independent values of significant wave height, a new declustering method was developed by Soukissian and Kalantzi (2009) that detects sequences of almost independent maxima from the initial time series in hand based on the physical features of a sea-state system. Specifically, large wave energy reductions between local maximum and subsequent minimum values of significant wave height imply the transition to a different sea-state system and hence leads to the identification of clusters of extreme events from the data series that are independent. After a simple filtering procedure of the initial time series using monotonicity for the detection and removal of stationary sequences, the local maxima and minima are identified and then the corresponding wave energy differences are calculated. If the wave energy reduction is lower than a predefined percentage, then it is considered that the examined sea-state system has ended forming thus a separate independent cluster. Again, the maximum value within each cluster is extracted to fit the GPD model. A rational selection of energy reduction percentage is over 80% that was also adopted in that work. The use of this declustering technique can be found in the studies of Soukissian and Arapi (2011).

3.7   Including directionality in extreme value estimation

Extremal properties of various environmental parameters have been modelled taking into account the directional behaviour as a covariate in order to obtain an integrated and more accurate model for the estimation of the corresponding design values. For instance, Moriarty and Templeton (1983) estimated extreme wind gusts for six directional sectors by fitting a GEV distribution in the design of large buildings. Maximum wind speed as a function of direction has also been modelled by Coles and Walshaw (1994), considering a dependence structure across directions, because their a priori division leads to correlated directional sectors, and adapting techniques developed for spatial extremes. Similar approaches for modelling extreme wind speed with a directional dependence structure have been presented by e.g. Simiu et al. (1985) and Solari and Losada (2016). A methodology for the appropriate selection of uncorrelated directional sectors has been proposed by Folgueras et al. (2019), which reduces also the uncertainty in the estimation of design values of wind speed. Sea currents have been investigated in the work of Robinson and Tawn (1997) by means of a parametric model for extreme current data by handling not only directionality but temporal dependence and non-stationarity as well.

In a series of papers, Ewans and Jonathan (2006, 2007, 2008) and Jonathan et al. (2008) have highlighted the importance of including directionality when studying extreme wave design criteria especially in storm-dominated regions. In the above studies, extreme value modelling of storm peak significant wave height was based on GP distribution with its unknown parameters expressed as a function of direction while a risk-cost approach was proposed for the construction of directional design criteria.

3.7.1   Extreme value directional model

Let a sample with values for a linear variable X along with the corresponding values for the directional one, say θ. Assuming that the GP distribution describes the extreme observations above a threshold u, which is considered independent of the directional variable, and according to Eq. (3.8), the cdf is given by

$$H_{Y_j|\theta_j.u}(y; \sigma_u, \xi) = 1 - \left(1 + \frac{\xi(\theta_j)y}{\sigma_u(\theta_j)}\right)^{-1/\xi(\theta_j)}, y > 0; \ \sigma_u > 0, \qquad (3.17)$$

for $1 + \xi(\theta_j)y/\sigma_u(\theta_j) > 0$, where shape and scale parameters are both expressed as functions of θ_j, with $\{\theta_j\}_{j=1}^{n_u}$.

In the context of estimating the unknown parameters, as noted by Robinson and Tawn (1997), it is expected that they vary smoothly with direction; thus, a Fourier series expansion is used for the description of this (angular) dependence, which assures a periodic behavior of the estimates in terms of the direction. In this respect, the general form of the Fourier series is for ξ and σ_u

$$\sum_{k=0}^{p}\sum_{b=1}^{2} A_{bk}t_b(k\theta) \text{ and } \sum_{k=0}^{p}\sum_{b=1}^{2} B_{bk}t_b(k\theta), \tag{3.18}$$

respectively, where $k = 0, \dots, p$ denotes the order of the Fourier model, and t_1, t_2 is the cosine and sine function, respectively. For example, the first order Fourier model results in the following relationships:

$$\xi(\theta) = A_{10} + A_{11}\cos(\theta) + A_{21}\sin(\theta) \text{ and } \sigma(\theta) = B_{10} + B_{11}\cos(\theta) + B_{21}\sin(\theta).$$

As noted by Jonathan and Ewans (2007), the proper order of the model is determined by the directional dependence of the data sample in hand; the more complex the directional dependence that characterize the data, the higher the model order is.

The unknown parameters A_{bk} and B_{bk}, $b = 1,2$, $k = 0, \dots, p$, are estimated by applying ML estimation. The likelihood of the corresponding data sample $\{Y_i\}_{i=1}^{n_u}$ is obtained by

$$L\left(\{A_{bk}\},\{B_{bk}\};\{Y_j\}_{j=1}^{n_u}\right) = \prod_{j=1}^{n_u}\frac{1}{\sigma_u(\theta_j)}\left(1 + \frac{\xi(\theta_j)}{\sigma_u(\theta_j)}Y_j\right)^{-(1/\xi(\theta_j))-1}, \tag{3.19}$$

and the negative log-likelihood (for $\xi(\theta_i) \neq 0$) by

$$\ell = \sum_{i=1}^{n_u}\left[\log\sigma_u(\theta_i) + \left(1 + \frac{1}{\xi(\theta_i)}\right)\log\left(1 + \frac{\xi(\theta_i)}{\sigma_u(\theta_i)}Y_i\right)\right]. \tag{3.20}$$

ML estimates can be determined by setting the partial derivatives of ℓ with respect to A_{bk} and B_{bk} set of parameters equal to zero, i.e.

$$\frac{\partial\ell}{\partial A_{bk}} = \sum_{j=1}^{n_u}\left\{\left[-\frac{1}{[\xi(\theta_j)]^2}\left(\log\left(1 + \frac{\xi(\theta_j)}{\sigma_u(\theta_j)}Y_j\right)\right.\right.\right.$$
$$\left.\left.\left. - \left(1 + \xi(\theta_j)\right)\left(\frac{\xi(\theta_j)Y_j}{\sigma_u(\theta_j) + \xi(\theta_j)Y_j}\right)\right)\right]t_b(k\theta_j)\right\} = 0$$

and

$$\frac{\partial\ell}{\partial B_{bk}} = \sum_{j=1}^{n_u}\left\{\frac{1}{\sigma_u(\theta_j)}\left[\frac{\sigma_u(\theta_j) - Y_j}{\sigma_u(\theta_j) + \xi(\theta_j)Y_j}\right]t_b(k\theta_j)\right\} = 0,$$

respectively.

The design values for a given return period can be estimated by using Eq. (3.13) and replacing the constant estimated parameters $\hat{\xi}$ and $\hat{\sigma}_u$ with the ones expressed as function of θ.

Penalised maximum likelihood

In this thesis, a penalty criterion is recommended for the extreme value estimates to ensure that that the directional dependence of ξ and σ_u is sufficiently described and that the solution is stable even if either the order of the Fourier model is high or the weighting constant of the penalty term is small, as is presented in Section 4.5. This penalty term is based on the absolute difference between the estimates and the initial values of the parameters obtained from the independent fits calculated using data from successive directional sectors of 45-degree width so that $\xi(\theta)$ and $\sigma_u(\theta)$ are consistent with ξ and σ_u obtained from the independent fits of each directional sector. As is discussed in Section 4.5, the minimum number of the 45-width sectors with sufficient amount of data should be set, which depends on the order of the Fourier model, along with the amount of data of each sector per se. With the inclusion of the penalty term in the model fitting, the terms that are not consistent are penalized appropriately. In this case, the negative log-likelihood with the penalty term takes the form

$$\ell_P = \ell + w \sum_{i=1}^{2(1+2k)} \left| \vartheta_i - \hat{\vartheta}_i \right|, \tag{3.21}$$

where w is a constant that gives the appropriate weight for the penalty term in model fitting and ϑ_i, $\hat{\vartheta}_i$ denote the initial and final values of the unknown parameters, respectively, with k indicating the order of the model. In Ewans and Jonathan (2008), a roughness penalty, selected using the cross-validation criterion, was adopted in order to obtain as smooth as possible estimates. In Figure 3-3, a preliminary result is presented for two locations, Ligurian and Aegean Sea (further analysed in Section 4.5), which shows the instability of a high order Fourier model. The solid lines denote the form of the estimated parameters ξ and σ_u obtained from the standard ML and the dashed lines denote the penalized version of ML with $w = 1$. These results clearly show the instability of the standard ML method when the order of the Fourier model is high. For these particular orders, the Fourier model has a better fit compared with the independent fits with data from eight consecutive sectors of 45-degree width while the standard directional model shows a rather oscillatory behaviour with a poor performance.

The directional extreme value model can be determined if the order of the model k is specified and the constant w is selected. In order to justify whether the inclusion of the directional covariates into the model is significant and judge which order of the Fourier model is the most adaptable in terms of capturing directional dependence, the likelihood-ratio (LR) test can be applied (Coles, 2001; Reiss and Thomas, 2007). This test is widely used when nested models are compared. Suppose that the basic model M_0 is nested within model M_1, which is more complex (e.g. the zeroth- and first-order directional models, respectively) with values of the negative log-likelihood ℓ_0 and ℓ_1, respectively. The LR test statistic is then expressed as

$$T_{\mathrm{LR}} = -2\big(\ell_0(M_0) - \ell_1(M_1)\big). \tag{3.22}$$

Under the null hypothesis that model M_0 is the true model, the distribution of T_{LR} is evaluated by assessing whether the additional complexity of model M_1 leads to a better improvement in terms of performance compared to model M_0. The asymptotic distribution of T_{LR} under the null model is a χ_k^2 distribution with k denoting the degrees of freedom equal to the difference among the number of the models parameters. As long as the sample size is reasonably large, it is

common to assume that this distribution is valid for finite samples as well. Consequently, the null hypothesis is rejected at the α level of significance if T_{LR} exceeds the $(1-\alpha)$ quantile of the χ_k^2 distribution. Hence, model M_1 is selected in favour of model M_0.

(a)   (b)
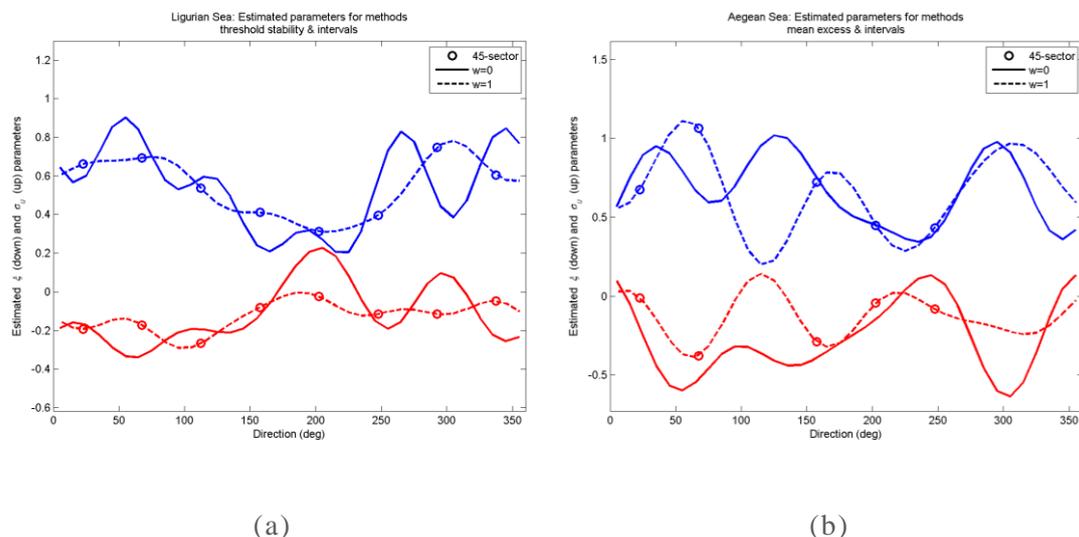
Figure 3-3. Estimated parameters ξ and σ_u for (a) a 5th and (b) a 4th order Fourier model with the consideration of the penalty term (dashed line) and without (solid line). Circles represent the estimates from the independent fits of the 45-degree sectors.

Given the order of the Fourier model for $\xi(\theta)$ and $\sigma_u(\theta)$, the constant w has to be set. This selection is based on the distance between the values of the estimated parameters ξ and σ_u from the independent fits and the corresponding ones from the directional model. The statistical metric that was selected due to the fair treatment of positive and negative differences is the mean absolute error. The optimum value for w is selected when the metric is minimized for both parameters simultaneously.

An example on the decision of the order of the Fourier model through the LR test and the selection of the constant w through the mean absolute error is provided for an offshore location in the Ligurian Sea with geographical coordinates (43.25°N, 9.75°E). The particular location (called hereafter Ligurian) was selected in order to have sufficient data for all directional sectors of width 45° and assess more reliable the directional model. The pairs of models used to perform the LR test are G_0 with G_1, G_1 with G_2, G_2 with G_3, and G_3 with G_4. The first pair is used in order to verify whether the use of directional model is rational for the data examined and the rest pairs to assess the performance of each order. The critical value corresponding to χ_k^2 for α=0.05, with which the T_{LR} is compared to, is the same for each case and equal to 9.4877, since the difference in the number of parameters remains the same (i.e. k=4) for each comparison. The obtained results are presented in Table 3-1 with all models being evaluated with w=1. Changes in w do not alter the result qualitatively. Based on the values of T_{LR}, compared to the critical one, with the smallest p −value, the third order directional model seems to be the most appropriate for Ligurian. Let us note that there is weak evidence to accept the null hypothesis for the comparison between G_0 and G_1, since the p −value is rather high denoting high uncertainty of the result.

In Figure 3-4, the functional forms of the estimated parameters for the first up to the forth order of the directional model is presented along with the corresponding estimates from the data obtained from successive directional sectors of width 45°. This outcome also verifies the above result; the third order model outperforms the first and second order models while the difference with the fourth model is rather unnoticeable. For w=0.18, the minimum value of mean absolute error for both estimated parameters is obtained as shown in Figure 3-5. The above results are

obtained for the 95th percentile as regards threshold selection and the intervals declustering method.

Table 3-1. Value of T_{LR} along with the corresponding $p-$value for various directional model comparisons for Ligurian Sea.

Pairs compared	$T_{\mathbf{LR}}$	$p-$value
G_0 and G_1	0.332	0.9877
G_1 and G_2	39.508	10^{-8}
G_2 and G_3	56.606	10^{-11}
G_3 and G_4	21.414	10^{-4}
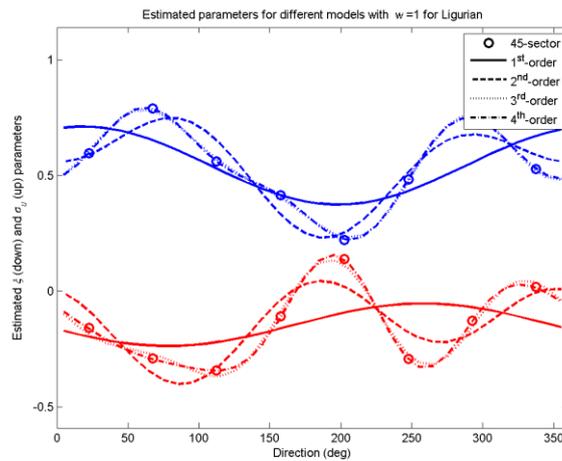
Figure 3-4. Directional model of 1st, 2nd, 3rd and 4th order along with the independent fits from the successive directional sectors of 45-degree width for Ligurian Sea.
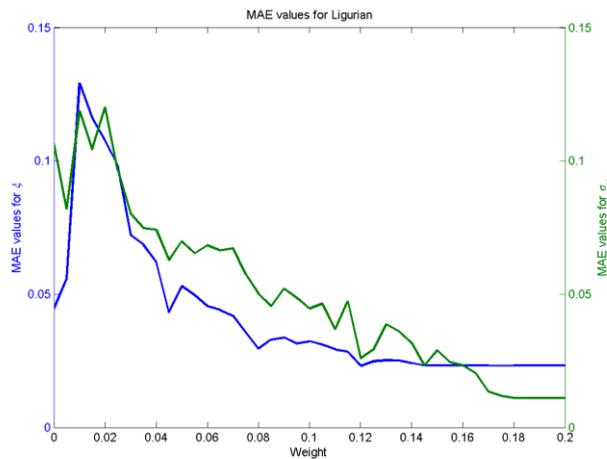
Figure 3-5. MAE values for various weights for Ligurian Sea.

3.7.2 Parameter uncertainty

Based on the expressions of the Fourier series for the estimation of ξ and σ_u, the corresponding asymptotic variances can be calculated so that confidence intervals for the unknown parameters

can be derived. The asymptotic covariance matrix of a ML estimator is provided by the inverse of the (observed) information matrix $I_{\hat{A}} = E\left[\partial^2 \ell / \partial A_{bk} \partial A_{\beta \kappa}\right]$ and $I_{\hat{B}} = E\left[\partial^2 \ell / \partial B_{bk} \partial B_{\beta \kappa}\right]$, which is actually the Hessian matrix (i.e. the matrix of the second derivatives of ℓ). The standard errors, $SE_{\hat{A}}$ and $SE_{\hat{B}}$, are the square roots of the diagonal elements of the asymptotic covariance matrix. Thus, the $100(1 - \alpha)\%$ confidence intervals of the parameters $\{A_{bk}\}_{b=1,k=0}^{2,p}$ and $\{B_{bk}\}_{b=1,k=0}^{2,p}$ are obtained, respectively, by:

$$\hat{A}_{bk} \pm z_{\alpha/2} SE_{\hat{A}} \text{ and } \hat{B}_{bk} \pm z_{\alpha/2} SE_{\hat{B}}, \tag{3.23}$$

where $z_{\alpha/2}$ is the desired critical value (e.g. for 95% confidence interval $z_{\alpha/2} = 1.96$). The above description is valid provided that the values of ξ fluctuate below $1/2$. Non-regular cases (i.e. for $\xi \geq 1/2$) are not considered in this thesis.

An alternative method for obtaining confidence intervals associated with the estimated parameters is through bootstrapping, introduced by Efron (1979) for samples collected under an independent framework. In extreme value analysis, typically bootstrap method is used in its non-parametric version (i.e. with no assumptions on the distribution of the available sample), and resample the original sample, while the parametric one simulates 'new' data from the estimated model); the former case is described in this section. Bootstrapping requires random resampling with replacement from the initial sample to obtain 'new' subsamples and construct confidence intervals for the parameters without assuming a specific parametric distribution, rendering the process quite straightforward, both algorithmically and numerically. Various bootstrap methods have been reviewed by Tajvidi (2003) for the construction of confidence intervals for the GP distribution parameters and quantiles and it was concluded that for small sample sizes none of the bootstrap methods gives satisfactory results. Moreover, Coles and Simiu (2003) proposed an empirical correction of the bootstrap estimates, based on a bias correction to the bootstrap parameter estimates, since there is a tendency of the bootstrap procedure to provide generally shorter tails than the one from the original time series. In this respect, the bias-corrected and accelerated (BCA) bootstrap method, developed by Efron (1987), is applied since it attempts to correct for both bias and skewness in the distribution of bootstrap estimates; for more details, see Efron and Tibshirani (1993).

Suppose that h is the parameter of interest and let us denote by $\hat{h}^*$ a bootstrap replication of $\hat{h}$ obtained by resampling with replacement from the original data sample. The underlying assumption of BCA method is that a monotone transformation $\phi = m(h)$ exists such that $\hat{\phi} \sim N(\phi - z_0(1 + a\phi), (1 + a\phi)^2)$, where z_0 and a are the bias-correction and acceleration constants, respectively. The former constant is related to the proportion of bootstrap estimates that are less than the corresponding estimate of the original sample and its estimate can be provided by

$$\hat{z}_0 = \Phi^{-1} \left\{ \frac{\widehat{\# \hat{h}^*}(r) < \hat{h}}{R} \right\}, \tag{3.24}$$

with Φ denoting the standard normal cumulative distribution function and $r = 1,2, \dots, R$ denoting each bootstrap sample with total number of bootstrap samples R. The latter correction is proportional to the skewness of the bootstrap distribution and can obtained by the jackknife method. Let $\hat{h}_{(i)}$, $i = 1, \dots, n$, denote the value of the estimate based on the entire original data sample apart from the $i-$th observation. An estimate of the acceleration constant is given by

$$\hat{a} = \frac{\sum_{i=1}^{n} (\hat{h}_{(\cdot)} - \hat{h}_{(i)})^3}{6 \left[\sum_{i=1}^{n} (\hat{h}_{(\cdot)} - \hat{h}_{(i)})^2 \right]^{1.5}}, \tag{3.25}$$

where $\hat{h}_{(\cdot)} = n^{-1} \sum_{i=1}^{n} \hat{h}_{(i)}$.

Having the values of $\hat{z}_0$ and $\hat{a}$, the interval of BCA method is given by $\left(\hat{h}^{(\alpha_1)}, \hat{h}^{(\alpha_2)} \right)$, where $\alpha_1 = \Phi \left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha)}}{1 - \hat{a}(\hat{z}_0 + z^{(\alpha)})} \right)$ and $\alpha_2 = \Phi \left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1 - \hat{a}(\hat{z}_0 + z^{(1-\alpha)})} \right)$ with $z^{(\alpha)}$ the 100α −th percentile point of a standard normal distribution.

Given the original (random) sample of pairs of one linear and one directional variable (x, θ), say $\{s_i\}_{i=1}^{n}$, the procedure of the adopted bootstrapping is summarized in the following steps for estimating the confidence intervals of the extreme value parameters:

Step 1: Estimate the unknown parameters $\left(\hat{\sigma}_u, \hat{\xi} \right)$ of the GP distribution (as functions of θ) from the initial sample using the ML method described above.

Step 2: Create r (random) samples $\left\{ s_i^{(r)} \right\}_{i=1}^{n}$, $r = 1, \dots, R$, by random resampling with replacement from the initial sample and obtain the estimates $\left(\hat{\sigma}_u^*, \hat{\xi}^* \right)$.

Step 3: Repeat step 2 for a large number R (of the order of 1000 or more).

Step 4: Estimate the two constants of BCA bootstrap method, $\hat{z}_0$ and $\hat{a}$ for each unknown parameter. Then estimate the lower and upper limits $\hat{\sigma}_u^{(\alpha_1)}$, $\hat{\xi}^{(\alpha_1)}$ and $\hat{\sigma}_u^{(\alpha_2)}$, $\hat{\xi}^{(\alpha_2)}$, respectively.

The same method can be applied to derive confidence intervals for return level x_T.

PART II

OFFSHORE & NEARSHORE APPLICATIONS

Chapter 4 Metocean climate modelling and analysis with applications in ocean energy assessment

4.1   General

Analysis and accurate description of metocean variables and their corresponding climatology is fundamental for many research purposes while it is involved in a variety of applied disciplines. Some of the most common cases include the assessment of marine energy resource, e.g. offshore wind (see, for example, Koletsis et al. (2016)) and wave (Cuttler et al., 2020), air and water pollution dispersion (Ramšak et al., 2013), weather routing along long-distance maritime routes (Perera and Soares, 2017), port infrastructure design and operation of offshore, nearshore and coastal structures (Sierra et al., 2017), coastal morphodynamics (Casas-Prat et al., 2016) and coastal zone management policies (Serafim et al., 2019) and marine ecosystems (Calvo et al., 2011). For such assessment studies, long-term data sets are required. The information contained in long-term metocean time series is important for assessing their variability, identifying potential correlations between two (or more) climatic variables, estimating extremes and design values, etc. Furthermore, the study of one metocean variable can shed light on the behaviour of other ones; for instance, atmospheric climate changes are likely to be reflected in the ocean surface physical characteristics (wind-driven circulation, Ekman transport) and the regional climate signal (e.g. water cycle, drought events, temporal variability) as well (Huang and McElroy, 2015).

As already mentioned, apart from the linear metocean characteristics, the corresponding directional features are also characterized by their inherent variability and their association structure with linear variables, thus they play an important role in many engineering and metocean studies. For example, in the context of ocean energy assessment studies which is the central core of applications presented in this chapter, directional parameters are involved, among others, in the micro-siting of offshore wind and wave farms (Song et al., 2016; Moarefdoost et al., 2017), the developing technology of floating wind turbines (Bachynski et al., 2014), the performance of wave energy converters for an accurate assessment of wave energy (Carballo et al., 2014) and fatigue analysis of offshore structures (Horn et al., 2018). Hence, the importance of analysing and modelling directional variables is evident, although the corresponding literature is either rather poor yet or often confined to the provision of standard rose plots.

This chapter consists of case studies coming from the fields of long-term climate modelling of metocean variables and marine renewable energy, and deals with real-world wind and wave data sets as described in Appendix E. Each section discourses topics related to the theoretical background of probabilistic modelling elaborated in Part I. Specifically, Section 4.2 presents an integrated approach for climate analysis and variability of wind speed and direction, Section 4.3 deals with the application of regression and calibration models for wind speed (under the presence of outliers) and wind direction data coming from different sources. Section 4.4 evaluates different bivariate models for the joint description of wind speed and direction and provides suggestions for the use of parametric and non-parametric models. Finally, Section 4.5 applies the directional extreme value analysis, a modification in the estimation of parameters of a directional extreme value model based on a penalised likelihood criterion is proposed and a thorough investigation of various methods of threshold selection and declustering is provided. At the time of presenting this thesis, the results presented in Sections 4.2, 4.3.2 and part of 4.4 have been published in three scientific journals, results from Section 4.3.3 have been published in the proceedings of a well-established international conference (International Ocean and Polar Engineering Conference) and from Section 4.5 have been submitted for publication.

4.2   Wind climate analysis and variability

4.2.1   Synopsis

In this section, the long-term offshore wind characteristics of the Mediterranean Sea are assessed based on a 36-year reanalyses data obtained by the ERA-Interim database. In order to identify the systematic wind flow patterns and reveal the general features of the wind climatology patterns, the wind climate analysis consists of the study of: i) the spatio-temporal behaviour (including variability characteristics) of wind speed and direction for the annual time scale; ii) the joint association of wind speed and direction for the annual time scale, and; the wind speed trends and wind direction changes. In the context of a climate assessment study, the analysis of wind direction changes are presented here for the first time.

The results of this analysis, along with other outcomes and discussion considering the monthly scale, have been published in:
Soukissian, T., Karathanasi, F., Axaopoulos, P., Voukouvalas, E.G., Kotroni, V., 2018. **Offshore wind climate analysis and variability in the Mediterranean Sea**. *International Journal of Climatology* 38: 384–402.

4.2.2   Short description of the study area description

The Mediterranean Sea, shown in Figure 4-1, is a semi-enclosed basin characterized by various geomorphological and topographical features with complex coastlines and local orography that influence the wind climatology both spatially and temporally. Some typical regional wind features occurring in the Mediterranean Sea are the following: Mistral, Tramontane, Bora, Sirocco, Etesian, Levante, Poniente, Leveche, etc. For a detailed description of the main Mediterranean winds, see Zecchetto and Cappa (2001) and references therein.

Apart from the aforementioned regional and local climatic features, the interaction of the airflow with the complex coastal orography, and the Mediterranean basin itself, plays also a significant role in the definition of the weather patterns, including the precipitating systems, the development of cyclones, etc. The Mediterranean Sea is one of the most cyclogenetic areas in the world (Flaounas et al., 2015), where explosive cyclogenesis (Lagouvardos et al., 2007) and tropical-like cyclones also occur (Tous and Romero, 2013).

Figure 4-1. The Mediterranean Sea divided into 11 sub-basins that are mentioned throughout the subsequent analysis (the background of the map has been derived from Google Earth).

4.2.3 Wind speed and wind direction climate

The notation of the following sections is introduced in Appendix B. The realizations of the linear variable X correspond to wind speed, denoted by u_i, and the corresponding realizations of the directional variable (i.e. wind direction) are denoted by θ_i. All the results refer to the annual time scale. The analysed data come from the ERA-Interim dataset with horizontal spatial resolution of approximately 80 km covering the period 1979–2014; for more details, see Appendix E.3.

The spatial distribution of the mean annual wind speed and wind direction are presented in Figure 4-2. The windiest areas of the Mediterranean Sea are the Gulf of Lion, the area surrounding the longitudinal axis of the Aegean Sea, the S Levantine Basin, the belt extending from the E Algerian Basin up to the Gulf of Gabes as well as the Alboran Sea. The overall highest mean annual wind speed is observed for the offshore area of the Gulf of Lion (7.4 m/s with corresponding mean wind direction 320.9°) and the second highest is observed for the central Aegean Sea (7.2 m/s with corresponding mean wind direction 4.0°). Regarding the wind directional patterns, they are fairly comparable with the results provided by Zecchetto and De Biasio (2007) for the period 2000–2004. Specifically, the analysis revealed many typical regional scale wind patterns such as the easterly Levanter wind in the Alboran Sea and the northeasterly (cold) Bora in the Adriatic Sea. The identification and quantification of these patterns contributes to the identification of ocean circulation patterns in the corresponding areas.

4.2.4 Association between wind speed and wind direction

In Figure 4-3, the spatial distribution of the linear–circular correlation coefficient $r_{u\theta}^2$ (see Appendix A.3 for the mathematical definition) between annual mean wind speed and annual mean wind direction is depicted. The highest values of $r_{u\theta}^2$ are observed across the eastern coasts of N Levantine Basin, the Gulf of Lion, the SE Alboran Sea, the SE Algerian Basin, the northern and eastern coasts of the Adriatic Sea, the W Balearic and S Tyrrhenian Seas, the southern part of the S Levantine Basin as well as the Aegean Sea, reaching values between 0.34 and 0.37. The analysis at the monthly temporal scale showed that there is a strong linear association between wind speed and direction in the Aegean Sea and the Gulf of Lion, mainly for June, July, August and September. For the Aegean Sea, this behaviour can be attributed to the prevalence of the Etesians that blow persistently and intensively mainly during these months.
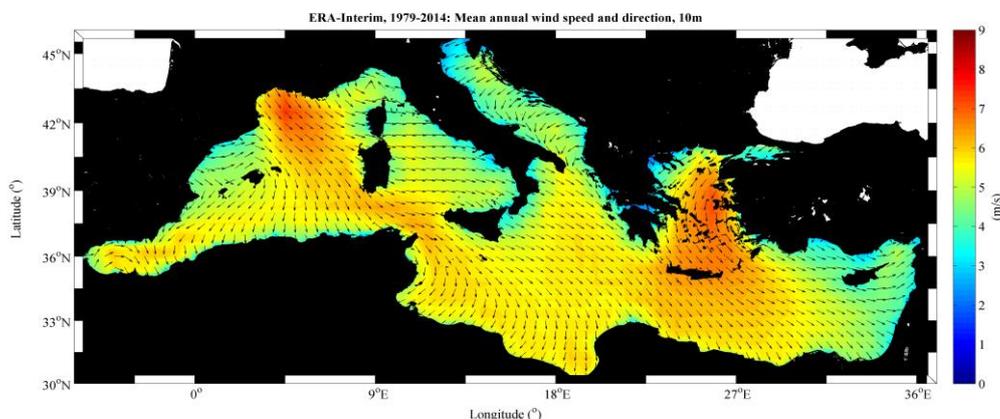
Figure 4-2. Spatial distribution of mean annual wind speed and wind direction over the Mediterranean Sea.
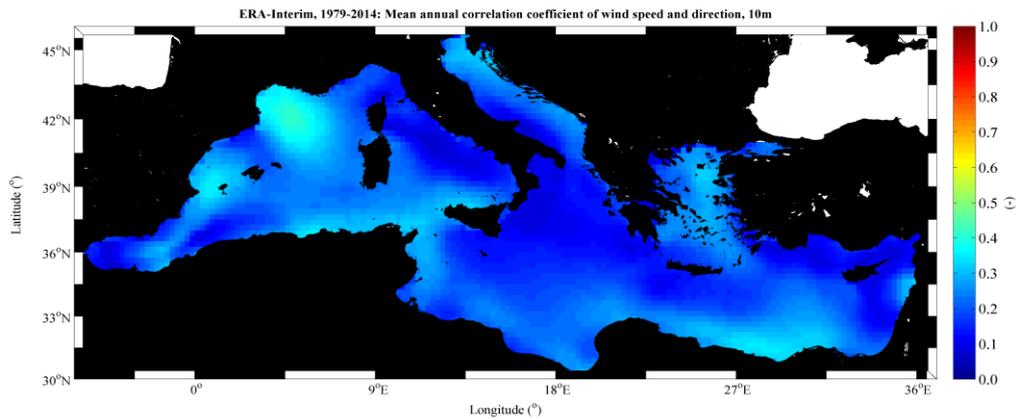
Figure 4-3. Spatial distribution of linear–circular correlation between wind speed and direction over the Mediterranean Sea on an annual basis.

4.2.5 Wind speed trend

The assessment of annual mean values for the estimation of linear slope for a seasonal series is supported by Withers and Nadarajah (2015). The authors suggest the use of annual mean values if data with duration equal to or longer than 5 years are available. The linear trend provides a quantification of the tendency of the mean intensity of wind fields in the examined time horizon. The parameter of main interest that quantifies the linear trend rate is the slope, i.e. the rate of wind speed change per unit time. In this analysis, the Theil-Sen linear estimation is adopted for the evaluation of the sought-for slope. This non-parametric estimator is less sensitive in departures of data from normality and less affected by the presence of outliers in the examined time series. In order to test the existence of a monotonic trend, the non-parametric Mann-Kendall test was adopted; see also Appendix B.6. The obtained results describe whether the statistical characteristics of wind speed tend to increase (positive values of the slope), decrease (negative values of the slope) or remain almost constant (slope close to zero) in the examined time scale.

In Figure 4-4, the spatial distribution of the linear slope of the annual mean wind speeds $m_{u,Y}(j), j = 1,2,..,J$, is depicted for the period 1979–2014. The dotted areas are characterized by statistically significant trends according to the Mann-Kendall test. The largest positive slopes are observed in the Ionian Sea (0.0159 m/s/year), the N Tyrrhenian and N Adriatic Seas, the eastern part of the Algerian Basin up to Balearic Isl. and the western part of the S Levantine Basin (between Crete Isl. and Africa). The overall minimum negative slope is observed offshore the coasts of Monaco in the Ligurian Sea (−0.023 m/s/year) while milder negative slopes appear also in the central Aegean Sea (−0.014 m/s/year), the E Alboran Sea and the N Levantine Basin. The long-term decrease of wind speed in the central Aegean Sea is in agreement with the results found by Poupkou et al. (2011).

The results from the wind speed trend analysis are in qualitative agreement with the ones from previous studies, although the spatial and temporal extent, and the source of the considered datasets may vary. For instance, Aarnes et al. (2015) investigated, among others, trends of global marine winds between 1979 and 2012 using also the ERA-Interim reanalysis dataset; the results for the trends in the Mediterranean Sea, presented in Figure 5(f) of Aarnes et al. (2015) are qualitatively in agreement with the ones presented in Figure 4-4.
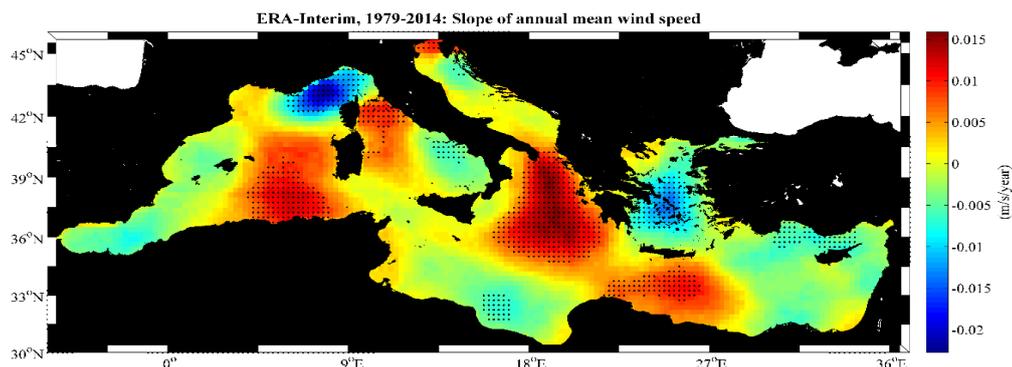
Figure 4-4. Spatial distribution of annual mean wind speed linear slope (in m/s/year) over the Mediterranean Sea. Dotted areas exhibit statistically significant trends.

4.2.6   Variability characteristics of wind speed and wind direction

The spatial distribution of mean annual variability is depicted in the upper panel of Figure 4-5. The highest variability within each year is exhibited in the northern part of the Adriatic Sea (68%), as well as the Ligurian, Tyrrhenian (especially offshore the northern coasts of Sicily), N Aegean and W Balearic Seas, and the Gulf of Antalya. According to Zecchetto and De Biasio (2007), the large wind speed variability in the Mediterranean Sea occurs in places swept by winds from the neighbouring land, like in the Adriatic Sea where Bora orographic downslope winds blow from Croatia, Slovenia and Montenegro, in the Ligurian Sea where winds blow from the coasts of France and Italy, over the N Aegean Sea and in the Gulf of Antalya. The spatial patterns depicted in this figure are in agreement with the patterns shown in Figure 3 of Zecchetto and De Biasio (2007).

In the lower panel Figure 4-5, the spatial distribution of inter-annual variability is shown. The strongest inter-annual signal appears in the W Ligurian Sea reaching values of the order of 7.25%. Other areas of relatively high inter-annual variability are the N Adriatic, Tyrrhenian and Balearic Seas, the Gulf of Lion, the S Algerian Basin and the Ionian and central Aegean Seas.

The angular variance of annual mean wind direction is depicted in Figure 4-6. In a large extent, the Mediterranean Sea is characterized by rather low values of angular variance fluctuating between 0.0 and 0.2. However, there are some areas that exhibit relatively high values (well above 0.4), namely the western part of the Alboran and Balearic Seas, the S Algerian Basin, the N Ligurian Sea and NW and S Adriatic Sea. The overall highest value of angular variance (0.975) is observed in the western part of Majorca Isl., and in the Ligurian Sea, eastern of Monaco (0.9) suggesting a highly fluctuating wind direction. Notice that for the area offshore the Gulf of Genoa, high variability of both wind speed and direction is observed.

ERA-Interim, 1979-2014: Mean annual variability of wind speed
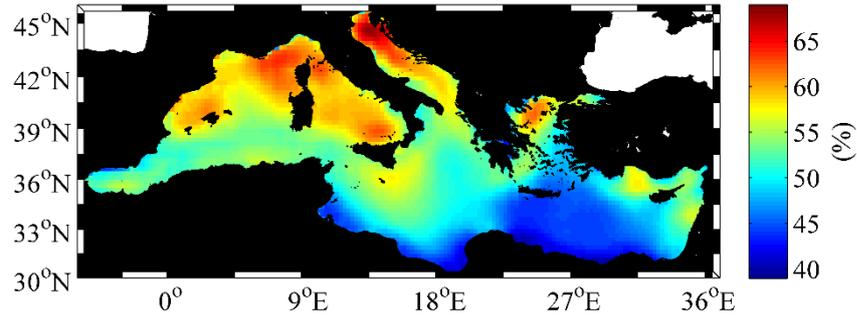
ERA-Interim, 1979-2014: Inter-annual variability of wind speed
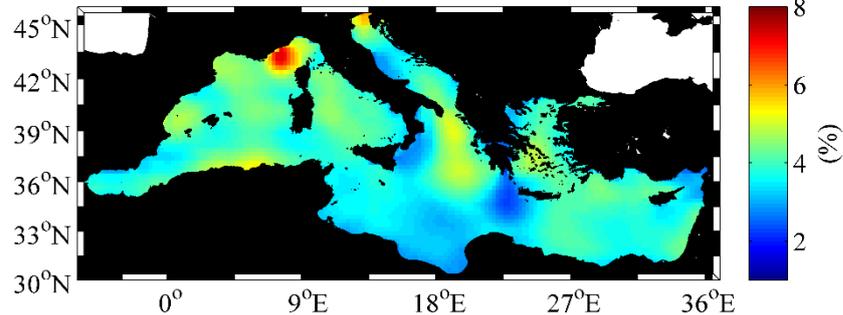
Figure 4-5. Spatial distribution of mean annual variability (upper panel) and inter-annual variability (lower panel) of wind speed over the Mediterranean Sea.

4.2.7 Wind speed and direction changes

In this subsection, the mean values of year-to-year angular distance of wind direction are provided. In order to quantify the wind direction changes, the year-to-year angular distance $\Delta\theta_{Y=j}$ is introduced, which is defined as:

ERA-Interim, 1979-2014: Angular variance of annual mean wind direction
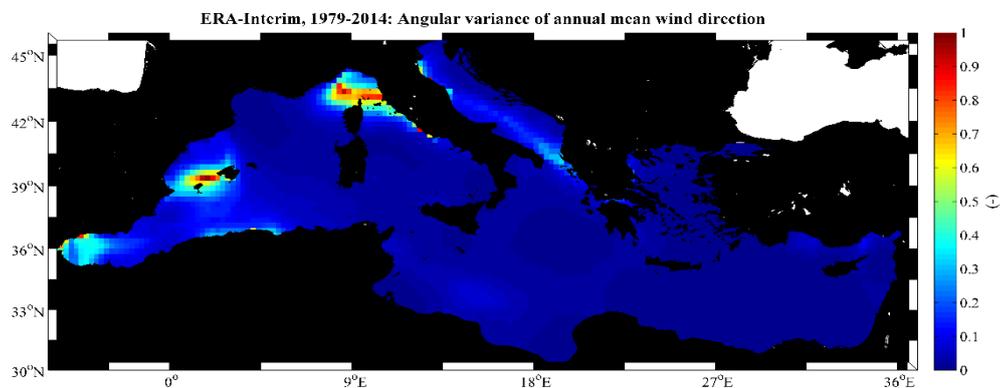
Figure 4-6. Spatial distribution of angular variance of annual mean wind direction over the Mediterranean Sea.

$$|\Delta\theta_{Y=j}| = \min(|\overline{\theta}_{Y=j+1} - \overline{\theta}_{Y=j}| \mathrm{mod} 2\pi, 2\pi - |\overline{\theta}_{Y=j+1} - \overline{\theta}_{Y=j}| \mathrm{mod} 2\pi), \qquad (4.1)$$

and bounded in $[-\pi, \pi]$, with

$$\begin{cases} \Delta\theta_{Y=j} \geq 0, & \text{if } \overline{\theta}_{Y=j} \leq \overline{\theta}_{Y=j+1} < \overline{\theta}_{Y=j} + \pi \\ \Delta\theta_{Y=j} < 0, & \text{if } \overline{\theta}_{Y=j} - \pi < \overline{\theta}_{Y=j+1} < \overline{\theta}_{Y=j} \text{ or } \overline{\theta}_{Y=j+1} > \overline{\theta}_{Y=j} + \pi \\ \Delta\theta_{Y=j} = (\overline{\theta}_{Y=j+1} - \overline{\theta}_{Y=j}) \mathrm{mod} 2\pi, & \text{if } \Delta\theta_{Y=j} = \pi, \end{cases} \qquad (4.2)$$

for $j = 1, 2, \ldots, J - 1$, where $\overline{\theta}_{Y=j}$ and $\overline{\theta}_{Y=j+1}$ denote the mean wind direction for years j and $j + 1$, respectively. $z \bmod w$ denotes the remainder when z is divided by w. Farrugia et al. (2009) provided a simpler expression for the estimation of $\Delta\theta_{Y=j}$, given by

$$\Delta\theta_{Y=j} = 2\tan^{-1}\{\tan[0.5(\overline{\theta}_{Y=j+1} - \overline{\theta}_{Y=j})]\}, \qquad (4.3)$$

Eq. (4.3) provides a signed value of $\Delta\theta_{Y=j}$, for a particular year j, indicating the direction of rotation, i.e. clockwise (positive sign) or anticlockwise (negative sign). In the same way, angular distances for other temporal scales can be defined. $\Delta\theta_{Y=j}$ is a random variable following a particular stochastic structure that, in the spatial domain, can be described through the corresponding mean value $\overline{\Delta\theta_Y}$.

In Figure 4-7, the spatial distribution of the mean year-to-year angular distance $\overline{\Delta\theta_Y}$ is presented, where clockwise (anticlockwise) rotation indicates areas with positive (negative) angular distance. In general, in most areas of the Mediterranean Sea, $\overline{\Delta\theta_Y}$ takes low absolute values. Areas of Balearic, Ligurian, Tyrrhenian and Alboran Seas exhibit the most evident extreme values of $\overline{\Delta\theta_Y}$. The overall extreme values (i.e. $-37°$ and $99.7°$) are encountered in the Ligurian/Tyrrhenian and Tyrrhenian Seas, respectively. In the Balearic Sea the corresponding values are $-35.8°$ and $16.8°$, respectively, while in the Alboran and Adriatic Seas the extreme values of $\overline{\Delta\theta_Y}$ are $-8.7°$ and $-9°$, respectively.

A comparison of the patterns from Figure 4-7 and Figure 4-6 reveals many similarities between them; areas characterized by high values of angular variance are also characterized by high values of angular distance. Comparing the results of Figure 4-4 and Figure 4-7, it is observed that some of the areas mentioned above exhibit simultaneously significant slopes of wind speed and year-to-year angular distances, namely the Ligurian and Tyrrhenian Seas. For instance, at the area around 43.25°N, 8.5°E, wind speed exhibits a long-term rate of change close to -0.016 m/s/year, while the corresponding year-to-year angular distance of wind direction is of the order of $-33°$.
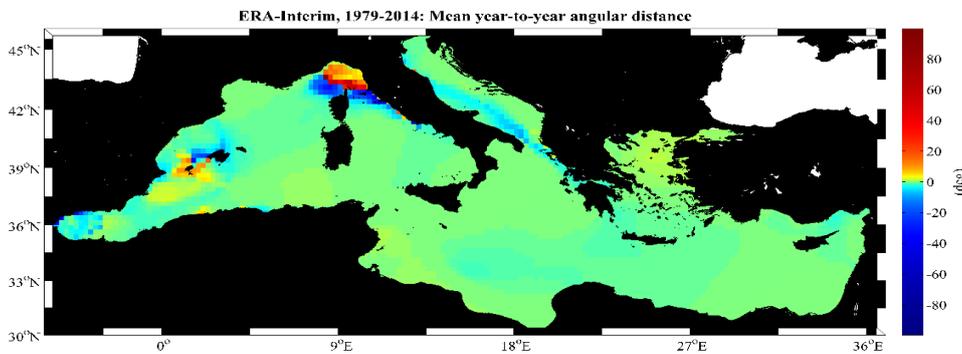
Figure 4-7. Spatial distribution of signed mean year-to-year angular distance over the Mediterranean Sea.

Another type of results presented here refers to the segmentation of the available time in time segments of 10 years, and the estimation of the relative change between the first and the last decade. Specifically, for wind speed the following quantity is estimated

$$\delta m_u^{I,II} = \frac{m_{u,Y_{II}} - m_{u,Y_I}}{m_{u,Y_{II}}}, \tag{4.4}$$

where m_{u,Y_I} denotes the mean annual value of wind speed obtained for 1979–1988, and $m_{u,Y_{II}}$ denotes the corresponding value for 2005–2014 in the results presented in this subsection.

Regarding wind direction, the angular difference between the mean annual values for the first and last decade of the time series is also provided. The angular difference is defined as follows:

$$\delta\theta^{I,II} = \min\left[\left|m_{\theta,Y_{II}} - m_{\theta,Y_I}\right|, 360 - \left|m_{\theta,Y_{II}} - m_{\theta,Y_I}\right|\right], \tag{4.5}$$

where $|\cdot|$ denotes the absolute value operator, m_{θ,Y_I} denotes the mean annual wind direction for the period 1979–1988, and $m_{\theta,Y_{II}}$ the corresponding value for the period 2005–2014. Eq. (4.7) provides the absolute difference between the mean annual wind directions corresponding to the examined decades, while $\delta\theta^{I,II}$ is bounded within [0,180]. $\delta\theta^{I,II}$ is a gross characteristic of the wind direction change between long-term time periods. Such information is valuable in the offshore wind energy industry and relevant fields of activity, while it could be also used as a potential indication of more significant environmental changes.

The spatial distribution of $\delta m_u^{I,II}$ is depicted in Figure 4-8. In order to secure the statistical validity of the results, the non-parametric Mann-Whitney U test has been applied. The null hypothesis that is tested is whether the examined samples (i.e. the mean annual values of wind speed from the first and the last decade) come from the same population. The use of the Mann-Whitney U test against the standard t-test is justified by the fact that the former test can be applied on unknown distributions in contrast to the latter test that can be applied only to samples from normal populations. In this figure, only the locations with critical p-values less than 0.05 are presented. For these areas, the p-values suggest the rejection of the assumption that the two samples come from the same population or have the same means. Therefrom, it can be concluded that, for the particular areas, the estimated relative wind speed changes are statistically significant. These areas are the N Adriatic Sea (Gulf of Venice, where the relevant increase of wind speed reaches values up to 10.2%), the S Ionian Sea (the relevant increase of wind speed reaches values up to 7.1%), the area offshore the Gulf of Genoa (the relevant decrease of wind speed reaches values up to –13.5%, which, in absolute terms, is the greatest change observed in the entire Mediterranean), the central Aegean Sea, particular areas in the N and E Tyrrhenian Sea, as well as some spots at the western and eastern part of the S Levantine Basin, part of the E Algerian Basin and the southern part of the central Mediterranean Sea, and the area between the northern coasts of Cyprus and the coasts of Turkey.

As already mentioned, studies on the long-term change of the wind flow over the whole Mediterranean are not available. The results presented here are in agreement with the results found specifically over the Aegean by Poupkou et al. (2011). Indeed the authors performed a trend analysis of the Etesian winds over the Aegean based on 31-year reanalyses data and they also found a negative trend in the frequency and wind speed of the Etesians.

In Figure 4-9, the spatial distribution of $\delta\theta^{I,II}$ is presented. Again, in order to secure the statistical validity of the results, the areas depicted in this figure are those with critical p−values of the Mardia-Watson-Wheeler test (see Appendix B) less than 0.05. The null hypothesis that is tested is whether the examined samples (i.e. the mean annual values of wind direction from the first and the last decade) have identical circular distributions regarding mean direction, circular variance or both. For these areas, the p−values suggest the rejection of the

assumption that the two samples come from the same population or have the same means. Therefore, it can be concluded that, for the areas depicted in Figure 4-9, the estimated angular differences are statistically significant.
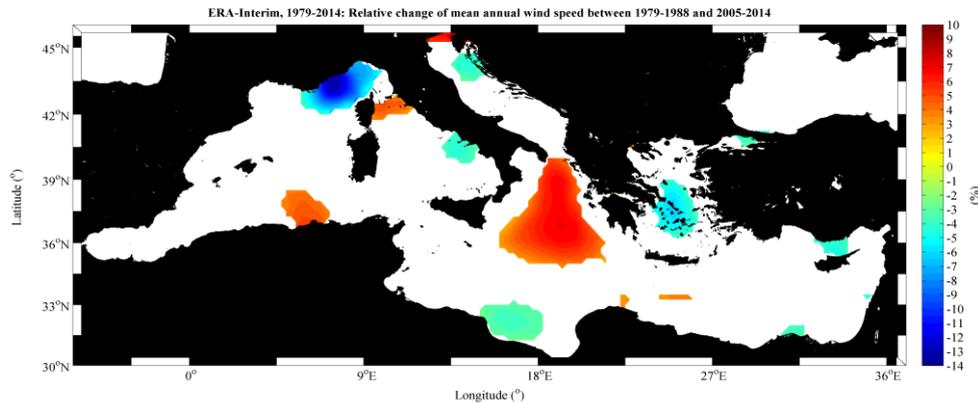
Figure 4-8. Spatial distribution of the relative change of mean annual wind speed between the first (1979–1988) and the last (2005–2014) decade of the available time series over the Mediterranean Sea. White colour denotes the areas where this relative change is not statistically significant based on Mann-Whitney U test.

According to Figure 4-9, the statistically most significant decadal angular differences occur in the northwestern part of the Adriatic Sea (angular differences up to 168°), the W Balearic Sea (angular differences up to 148°), and across the longitudinal axis of the W Adriatic Sea (around 70°). Milder differences (of the order of 10° to 30°) are observed for the extended area between the eastern part of the central Mediterranean Sea and western part of S Levantine Basin, NE Aegean Sea, N and NE of the S Levantine Basin, specific areas around S Italy, W Algerian Basin, and some spots north of Corsica (Ligurian Sea) and south of Sardinia (Tyrrhenian Sea). Let us note that although the magnitude of the relative change of mean annual wind direction for the examined decades is of primary concern, it seems that there is a systematic wind direction shift from NE to S in the Balearic Sea, while the corresponding shift in the Adriatic Sea is not unidirectional. Furthermore, an analysis between the first (1979–1988) and the right next decade (1989–1998), that is not presented here, reveals two localized areas that are not present during the examined decades: W Alboran Sea, with relative changes up to 80°, and SE N Levantine Basin, with relative changes up to 15°.
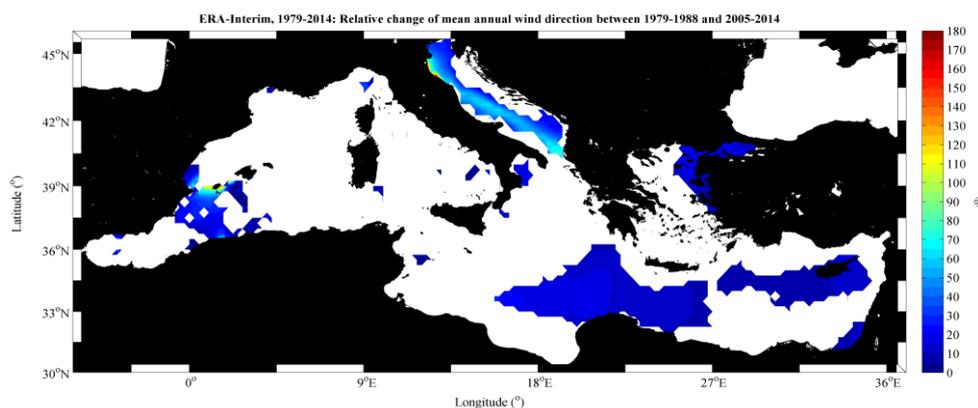
Figure 4-9. Spatial distribution of the angular change of mean annual wind direction between the first (1979–1988) and the last (2005–2014) decade of the available time series over the Mediterranean Sea. White colour denotes the areas where this relative change is not statistically significant based on Mardia-Watson-Wheeler test.

Overlapping the results of Figure 4-8 and Figure 4-9, it is observed that some particular areas exhibit statistically significant decadal changes both in wind speed and direction, namely a northern part of the Adriatic Sea, a southern part of the Ionian Sea, western of the Sea of Marmara, the western part of the S Levantine Basin (extending across the 33rd parallel North), a region offshore the Gulf of Genoa (Ligurian Sea) and north of Cyprus Isl. Summing up, the above results suggest that part of the observed wind variability in the specific areas could be attributed to the long-term change of wind direction patterns and not only on wind speed changes. This behaviour has also effects on wave propagation patterns and sea state intensities, since it also alters the corresponding fetch lengths and probably fetch durations. However, it would be risky to provide an explanation for the particular behaviour of wind direction in these areas due to the complex interactions and nature of the climatic system. Potential global warming effects could be one reason, since random fluctuations of wind direction, except for wind speed, may be caused for small-scale regions over short-term time scales.

4.2.8   Final comments

This study provides an overview of the linear and directional wind climate for the Mediterranean Sea, providing also some additional statistical characteristics such as wind variabilities, angular distances, and inter-decadal directional changes extending over the entire spatial scale of the examined basin, for the first time. Summarizing the major findings of this work, the following general conclusions can be drawn:

- Regarding wind speed variability, the areas that exhibit high values of mean annual and inter-annual variability are the N Adriatic, Tyrrhenian, Ligurian, Balearic, Ionian and Aegean Seas.
- Regarding wind direction variability, the W Mediterranean Sea (Ligurian, Tyrrhenian, Balearic and Alboran Seas) is characterized by high values of variance unlike the rest part of the basin.
- In terms of wind speed trend, positive values of slope throughout the examined period are observed in the Ionian, N Tyrrhenian and N Adriatic Seas, the eastern part of the Algerian Basin and the offshore area between Crete Isl. and Africa; large negative values are evident in the Ligurian and central Aegean Seas.
- In terms of angular distance, most areas of the Mediterranean Sea are characterized by low absolute mean values except for the Balearic, Ligurian and Tyrrhenian Seas.
- Regarding inter-decadal changes of wind speed, the highest positive decadal difference corresponds to the northern part of the Adriatic Sea (up to 10.2%) followed by the southern part of the Ionian Sea (up to 7.1%), while the highest negative decadal difference corresponds to the area offshore the Gulf of Genoa (up to −13.5%).
- Regarding inter-decadal changes of wind direction, the most significant directional differences are located in the northwestern part of the Adriatic Sea (up to 168°), the W Balearic Sea (up to 148°), and across the longitudinal axis of the W Adriatic Sea (around 70°).
- There is a simultaneous long-term large change of both mean annual wind speed and direction during the examined decades in a northern part of the Adriatic Sea, a southern part of the Ionian Sea, western of the Sea of Marmara, the western part of the S Levantine Basin (extending across the 33rd parallel North), a region offshore the Gulf of Genoa (Ligurian Sea) and north of Cyprus Isl.

Since wind speed trends and angular distances are evident, it is rational to assume that the underlying variations may be attributed to the distribution of high wind speeds through the changes in the occurrence of high impact weather related with wind storms, that is difficult to quantify; however, the reason behind these changes cannot be determined accurately in the context of this analysis nor the involved physical processes. On the other hand, large angular

distances for wind direction are expected to affect wave propagation schemes and the interdependent environmental phenomena. Finally, a longer-term data set could highlight whether the estimated variabilities and trends have a steady behaviour or constitute a part of a larger cyclic variation, as well as it could serve for a more robust analysis as regards the decadal changes of both wind speed and direction. Moreover, the results of this section clearly suggest that in any attempt for wind and wave climate analysis, directional characteristics should necessarily be taken into consideration.

4.3   Calibration of metocean characteristics

4.3.1   Synopsis

Buoy measurements are usually considered as a reference source in applications related to metocean climate analysis and site selection for ocean energy development including evaluation and calibration of metocean data obtained from less reliable sources, combined assessment, blending and homogenization of multisource metocean data, etc. Regarding linear variables, most of these applications are based on regression techniques elaborated by using the principle of ordinary least squares (OLS). However, long-term metocean data usually contain several outliers, which may question the validity of the regression analysis, if not properly considered. In this section, robust regression methods are implemented to identify and reveal outliers from wind data, and retain at the same time their efficiency. Long-term reference wind data series obtained from buoys at four locations in the Mediterranean Sea are used to calibrate hindcast (model) wind data by applying robust methods and OLS. The obtained results are compared according to several statistical measures. The effects of the calibration methods are also assessed with respect to the available wind power potential. The results clearly suggest that particular robust methods perform in all respects better than OLS.

On the other hand, calibration techniques are very rarely adopted for circular variables, although their accurate prediction seem to be significant in various applications. For instance, wind direction is a critical variable as regards the micro-siting procedure of offshore wind turbines within an offshore wind farm, since wake effects can affect the efficiency of the optimal aligning of turbines to wind direction (Castellani et al., 2015). Moreover, wave direction is also critical as regards the wave resource evaluation in an area when a wave energy converter is to be installed (Hiles et al., 2016). In this respect, wind (wave) direction not only should never be neglected in relevant applications but it should be determined as accurately as possible. In reality, the relevant scientific literature as regards calibration of wind and wave direction from various data sources implemented through linear regression analysis is rather poor. In this context, calibration techniques are presented for correcting wind direction at various locations in the Mediterranean Sea. The application data are measurements from in situ devices and results from NWP models and remote sensing instruments, which are corrected since they are considered to be less accurate. The obtained results suggest that the proposed statistical procedure should be applied along with the calibration of wind speed, whenever accurate data are required in wind energy assessments.

The results of the calibration of linear variables based on robust regression methods, along with outcomes from additional buoy locations, have been published in:
Soukissian, T.H., Karathanasi, F.E., 2016. **On the use of robust regression methods in wind speed assessment**. *Renewable Energy* 99: 1287–1298.

A similar study can also be found in:
Soukissian, T., Karathanasi, F. Voukouvalas, E., 2014. **Effect of outliers in wind speed assessment**. Proceedings of the 24th *International Offshore (Ocean) and Polar Engineering* Conference, 1: 362-369, Busan, June 15–20.

Part of the results of the calibration of directional variables, have been published in:
Karathanasi, F.E., Soukissian, T.H., Axaopoulos, P.G., 2016. **Calibration of wind directions in the Mediterranean Sea**. Proceedings of the *26th International Ocean and Polar Engineering Conference*, 1: 491-497, Rhodes, Greece, June 26–July 1.

4.3.2   Calibration of linear data using robust regression methods

The main aim of the proposed methodology is to calibrate concurrent wind speed data from the less accurate source (NWP model results) using buoy data (which is the reference data source), through the implementation of a linear calibration procedure. Furthermore, the performance of the examined regression (calibration) methods is also assessed. Along with the OLS method, the robust methods that have been applied and examined in this analysis are the *MM* – estimation (MM), Huber's M–estimation (M–H), least trimmed squares (LTS) and *L₁ – estimate* (L₁).

Henceforth, let u_M denote the wind speeds obtained from the NWP model and u_B denote the wind speeds obtained from buoy measurements. $\hat{u}_M$ denotes the corrected (calibrated) wind speed from the NWP model. Although the regression parameters are firstly estimated by the abovementioned methods, hereinafter, calibration parameters and procedures are only addressed, implying that the corresponding regression parameters have been firstly estimated. Based on the classical calibration model, let $\widehat{b}'_0$ and $\widehat{b}'_1$ denote the estimates of intercept and slope, respectively, of the calibration parameters.

Two different data sets are elaborated throughout the analysis: the first data set (called estimation set) is the one from which the calibration parameters are estimated, and the second data set (called evaluation set), is the one on which the calibration is applied and the evaluation of the methods is performed. See also Figure 4-10, for a schematic representation of the applied methodology. Specifically, the estimation dataset consists in collocated data from buoy measurements (black solid line) and NWP model results (black dashed line). From this data set, the parameters $\widehat{b}'_0$ and $\widehat{b}'_1$ for the calibration of the NWP model results are estimated. These parameters are used for acquiring the calibrated NWP model results (red solid line) of the evaluation data set and then the calibrated NWP model data are compared with the buoy data of the evaluation set.

The following special cases are examined in detail:

C.1 <u>Full data sample analysis</u>: This is the most fundamental case, where the estimation and the evaluation data sets are the same and refer to the entire available time period. Specifically, OLS and robust methods are applied using the full concurrent available wind data sets in order to calibrate (correct) wind speed data from the less accurate source. The evaluation of the performance of OLS and robust methods is made using the same data set.

C.2 <u>Partial data sample analysis</u>: The estimation and the evaluation data sets are different and non-overlapping, i.e. the estimation of the calibration parameters is made using part of the available data sample and the calibration and evaluation of the methods is made based on the remaining part of the available data sample. Specifically, the estimation data set refers to wind data corresponding to the first year of the available time period and the evaluation data set refers to data corresponding to the remaining time period. It is evident that the unused wind data sample is considered to contain new ("fresh") wind data, rendering this sample more realistic and the methodology tighter for the assessment of the calibration relations. This concept is closely related to the measure–correlate–predict (MCP) family of methods used in wind speed forecasting to extrapolate and extend in time short-term

wind time series; see, for example, Lackner et al. (2008) and a review from Carta et al. (2013).
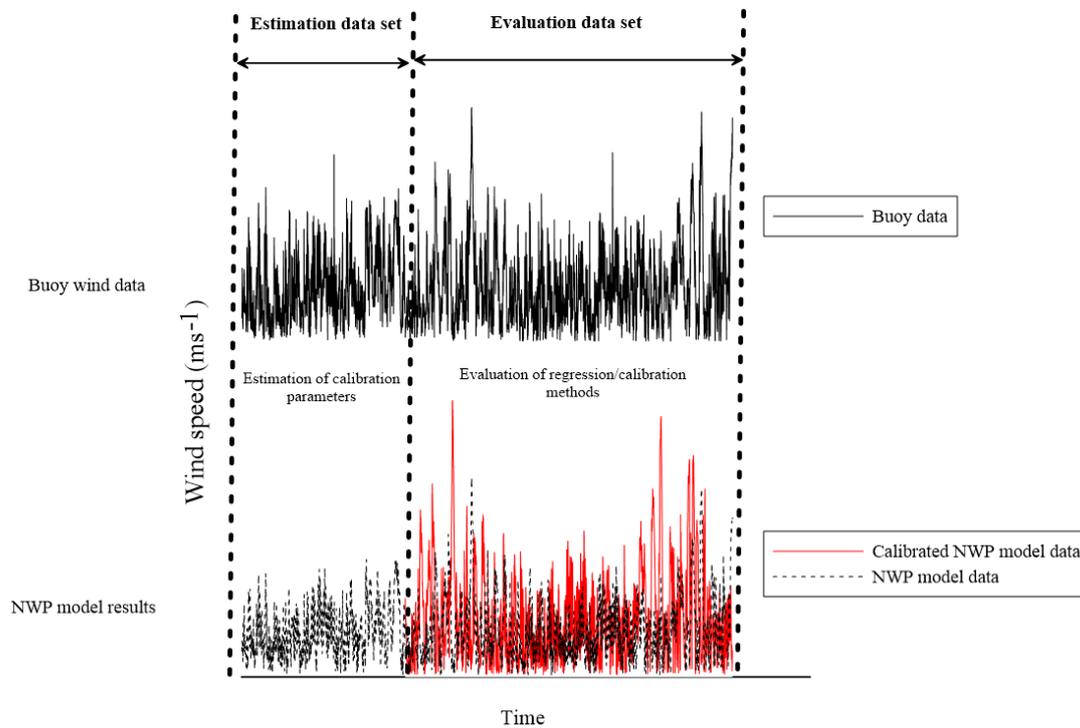
Figure 4-10. Schematic representation of the applied methodology corresponding to case C.2.

Two additional cases have also been examined in the work of Soukissian et al. (2014). In the first one, the calibration parameters were estimated from a time period in the middle of the recording period of the obtained datasets in order to correct and evaluate the less accurate data source from the time periods before and after the estimation set. In the second one, outliers were omitted from the full data sample based on the severity of change they cause on the regression parameters $\hat{b}_0$ and $\hat{b}_1$ in order to obtain a 'clean' dataset for OLS estimation; then, the OLS calibration parameters were re-estimated for the entire available time period, the data of the same period were re-corrected and the method was re-evaluated.

Overall, the general proposed methodology consists of the following steps:

S.1 Estimation of the calibration parameters $\hat{b}'_0$ and $\hat{b}'_1$ by using OLS and robust techniques from the collocated data samples referring to the estimation sets of C.1 and C.2 (the general recommendation is to use concurrent data of, at least, one year length due to seasonal wind variations).

S.2 Based on the parameters obtained from step S.1, correction of the wind speeds of the less accurate data source (NWP model data) referring to the evaluation sets of C.1 and C.2.

S.3 Comparison between the calibrated NWP model wind speeds (from step S.2) and the measured wind speeds obtained from buoys for the evaluation sets of C.1 and C.2.

In order to evaluate the performance of the examined regression/calibration methods, the calibrated NWP model wind speeds from robust methods and OLS are compared to the corresponding measured wind speeds obtained from the buoys through the following statistical measures: *BIAS*, *RMSE*, *MAE* and *SI*; for the definitions, see Appendix B.4.

Data and numerical results

Wind measurements from two buoys deployed in the Aegean Sea and two buoys in the Spanish waters are used as reference data source. Their geographic coordinates along with the examined measurement periods are presented in Table 4-1.

Before the regression/calibration analysis, the buoy wind data were first checked qualitatively and any missing or clearly erroneous values (such as spikes) were discarded. Then, for comparison purposes, wind speeds were adjusted to the reference level of 10 m above sea level using the log-law wind profile. After this adjustment, the collocation in space and time procedure was carried out. For the spatial collocation, the four nearest wind data series of the NWP model were downscaled to the exact location of each buoy by applying the weighting interpolation scheme. Regarding the temporal collocation, the common time frame was 3 h (00:00, 03:00, 06:00, etc. UTC).

In Table 4-2, the results of a primary statistical analysis regarding the concurrent wind datasets (buoy measurements and NWP model results) are presented for the examined locations. The statistical parameters depicted in this table are the following: sample size N, mean value m, standard deviation s, minimum value min, maximum value max, coefficient of variation CV and coefficient of determination r^2 between buoy measurements and NWP model results. It should be noted that, despite the collocation procedure, the sample size remains adequate for performing a statistically reliable analysis.

Estimation of regression/calibration parameters

The regression lines obtained from all examined methods (OLS method, MM-estimation, Huber's M-estimation, LTS and L1-estimation) for the entire samples (i.e. case C.1) along with the corresponding density scatter plot are shown in Figure 4-11. The colour gradation indicates the density (percentage) of the data points falling within each square, where the red and blue tone denotes the maximum and minimum frequency of appearance, respectively. In general, the OLS line appears to be relatively further off the diagonal compared to the lines obtained from robust methods. In addition, the slope of the regression line $\hat{b}_1$ is systematically lower for OLS compared to robust methods in all examined cases indicating the lower variance of the predicted values of model data. It is also easily observed that robust methods provided values of slope and intercept mutually close, compared to the values provided by OLS.

Table 4-1. Names, geographical coordinates and concurrent measurement time periods for the examined buoys.

Buoy name	Location (lat, lon)	Recording period
Lesvos	39°10′N, 25°49′E	1/2000−12/2004
Mykonos	37°31′N, 25°28′E	1/2000−12/2004
Cabo Begur	41°55′N, 3°39′E	3/2001−12/2004
Cabo de Gata	36°43′N, −2°19′E	1/2000−10/2004

Table 4-2. Basic statistics of wind speed for all the examined sites in the Mediterranean Sea for the concurrent recording periods.

Location	Data source	N	*m* (m/s)	*s* (m/s)	min (m/s)	max (m/s)	CV (%)	r^2
Lesvos	Buoy	11453	7.14	4.07	0.12	34.42	57.07	0.59
	Model		5.75	3.05	0.10	22.92	53.11	
Mykonos	Buoy	8606	8.14	4.18	0.12	21.48	51.33	0.71
	Model		6.62	3.21	0.16	17.97	48.46	
Cabo Begur	Buoy	2257	8.22	5.85	0.23	27.72	71.17	0.74
	Model		6.70	3.97	0.18	22.50	59.20	
Cabo de Gata	Buoy	8870	6.20	3.83	0.23	20.99	61.76	0.70
	Model		5.40	3.38	0.10	18.38	62.50	

(a)

(b)

(c)

(d)
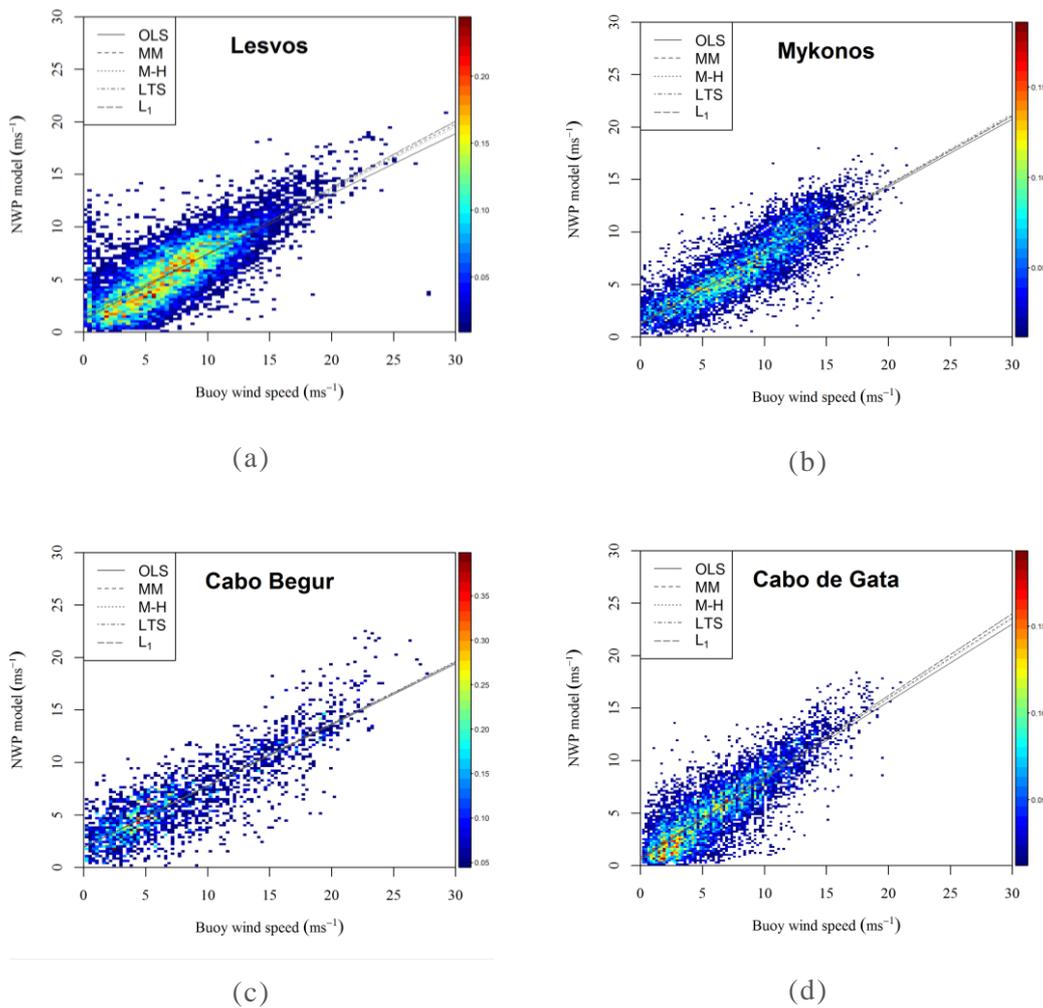
Figure 4-11. Scatter plot (with colour-scale indicating density) and regression lines of wind speeds obtained for case C.1 for all the applied methods at: (a) Lesvos, (b) Mykonos, (c) Cabo Begur and (d) Cabo de Gata.

The corresponding calibration parameters $\widehat{b}'_0$ and $\widehat{b}'_1$ for the NWP model wind data are estimated. The values of these estimated parameters, for each method and both cases, are presented in Table 4-3. From this table the following conclusions can be drawn:

i. Parameter $\widehat{b}'_0$ is always negative and parameter $\widehat{b}'_1$ is always positive for all methods, locations and cases examined. This behaviour is due to the fact that the NWP model tends to underestimate wind speed measured from buoys, as was mentioned in Section 0;

ii. the estimated parameters from the pairs of MM and M–H methods, and LTS and L_1 methods are fairly close to each other in the majority of the locations and cases examined;

iii. the values of parameters $\widehat{b}'_0$ and $\widehat{b}'_1$ of the OLS method are consistently lower and higher, respectively, compared to the corresponding parameters of the robust methods. Exception of this behaviour is $\widehat{b}'_0$ in Cabo Begur for case C.1;

iv. in general, LTS and L_1 methods provide the smallest values $\widehat{b}'_1$ parameter for the majority of the examined locations.

Evaluation of calibration methods

The evaluation of the performance of the examined calibration methods is made by estimating the statistical measures mentioned before. In case C.1, the estimation of the calibration parameters is made by utilizing the entire concurrent data samples of buoy measurements and NWP model results. In Table 4-4, the obtained values of the applied statistics are presented at the examined locations for OLS and those of the examined regression/calibration methods that provided at least one minimum value of any of the statistics used. The optimum value of the statistics for each examined case is shown in boldface letters. The most important conclusions that can be drawn from the obtained results are the following:

- MAE and RMSE (along with SI), which are statistic measures quantifying the absolute and squared difference between corrected and measured wind speeds, respectively, were systematically lower for LTS and L_1 methods for all examined locations.
- From the overall combination of four locations with four statistical criteria (i.e. in total 16 outcomes), LTS performed better for 12 out of the 16 outcomes and L_1-estimator for 5 outcomes (in *SI* results of Gabo de Gata both methods performed equally best).

Table 4-3. Estimated calibration parameters obtained from each applied method for cases C.1 and C.2 for all examined locations.

Location	Method	$\widehat{b}'_0$		$\widehat{b}'_1$	
		C.1	C.2	C.1	C.2
Lesvos	OLS	−2.865	−2.087	1.740	1.657
	MM	−2.059	−1.771	1.622	1.604
	M–H	−2.241	−1.804	1.649	1.610
	LTS	−1.806	−1.634	1.587	1.581
	L_1	−1.849	−1.551	1.589	1.559
Mykonos	OLS	−2.113	−2.753	1.548	1.666
	MM	−1.867	−2.497	1.511	1.627
	M–H	−1.900	−2.539	1.517	1.635
	LTS	−1.750	−2.349	1.494	1.608
	L_1	−1.888	−2.473	1.516	1.631
Cabo Begur	OLS	−3.285	−3.338	1.715	1.776
	MM	−3.273	−3.217	1.703	1.731
	M–H	−3.279	−3.262	1.705	1.743
	LTS	−3.300	−3.177	1.697	1.712
	L_1	−3.231	−3.289	1.701	1.743
Cabo de Gata	OLS	−1.104	−0.971	1.353	1.307
	MM	−0.837	−0.738	1.304	1.265
	M–H	−0.883	−0.774	1.313	1.272
	LTS	−0.700	−0.646	1.280	1.248
	L_1	−0.676	−0.579	1.278	1.242

Table 4-4. Statistics of calibration equations based on OLS, MM, M-H, LTS and L_1 for all the examined locations for case C.1.

Location	Method	BIAS	RMSE	MAE	SI
Lesvos	OLS	0.347	3.317	2.501	0.441
	LTS	0.338	**3.107**	**2.323**	0.424
	L_1	**0.315**	3.108	2.325	**0.423**
Mykonos	OLS	0.078	2.664	2.064	0.322
	LTS	**0.062**	**2.583**	**2.000**	**0.313**
Cabo Begur	OLS	0.292	3.389	2.576	0.394
	LTS	**0.150**	3.350	2.545	0.389
Cabo de Gata	OLS	0.135	2.455	1.872	0.383
	LTS	**0.071**	2.347	1.783	0.373
	L_1	0.082	**2.345**	**1.781**	**0.373**

In case C.2, the estimation of the calibration parameters was made by utilizing only the first year data from the concurrent data samples of buoy measurements and NWP model results. After applying the calibration, the evaluation of the obtained results was made on the remaining time period data. In Table 4-5, the obtained values of the applied statistics are presented at the examined locations for OLS and those of the examined regression/calibration methods that provided at least one minimum value of any of the statistics used. Again, the optimum value of the statistics for each examined case is shown in boldface letters. The most important conclusions that can be drawn from the obtained results are the following:

- The best values for RMSE, MAE and SI are obtained from LTS and L_1-estimator robust methods for all the examined locations.
- Each of LTS and L_1-estimator methods performed better 9 and 6 times, respectively, while 1 time LTS and MM-estimator performed equally best. In total, together LTS and L_1-estimator performed better in 15 out of 16 total outcomes. Furthermore, OLS performed well only 1 time.

Evaluation of calibration methods on wind energy estimation

In this section, the effects that the different regression (calibration) methods have on the estimation of the mean wind power density $\bar{P}$ are assessed at the examined locations in the Greek and Spanish waters. The mean (long-term) wind power density $\bar{P}$ in a specific sea area can be directly obtained, if a sufficiently long time series of observed wind speeds is available, through the following relation:

Table 4-5. Statistics of calibration equations based on OLS, MM, M-H, LTS and L_1 for all the examined locations for case C.2.

Location	Method	BIAS	RMSE	MAE	SI
Lesvos	OLS	0.587	3.305	2.451	0.449
	L_1	**0.475**	**3.146**	**2.320**	**0.433**
Mykonos	OLS	0.347	2.865	2.226	0.350
	LTS	**0.325**	**2.758**	**2.141**	**0.339**
Cabo Begur	OLS	0.812	3.542	2.715	0.441
	LTS	**0.529**	3.363	2.577	0.419
Cabo de Gata	OLS	**−0.061**	2.358	1.786	0.379
	LTS	−0.099	2.283	1.726	**0.372**
	L_1	−0.083	**2.279**	**1.723**	0.373

$$\bar{P} = \frac{1}{2N} \sum_{i=1}^{N} \rho u_i^3, \tag{4.6}$$

where N is the sample size and u_i, $i = 1, ..., N$, is the observed wind speed time series. $\bar{P}$ can be directly estimated by utilizing the wind speed time series u_{B_i}, $i = 1, ..., N$, obtained from buoys. This estimate is denoted by $\bar{P}_B$ and is considered as the reference value. Moreover, $\bar{P}$ can be also estimated by using the calibrated results of the NWP model, i.e. the wind speed time series $\hat{u}_{M_i}$, $i = 1, ..., N$; this estimate is denoted by $\bar{\hat{P}}_M$. The quality of the calibration procedure described in the foregoing sections can be additionally cross-examined by evaluating the (absolute) relative error between $\bar{P}_B$ and $\bar{\hat{P}}_M$.

The obtained results for relative errors with respect to all locations and regression/ calibration methods examined are shown in Table 4-6. The minimum values of this quantity for each examined case and method is shown in boldface letters. It is clear from these results that LTS and L_1-estimation methods perform systematically better than OLS. It is also worth mentioning that OLS method provides the largest relative error for all cases and locations. Let it be noted that, for case C.1, the relative error obtained for the non-calibrated data is significantly reduced by all calibration methods (results are not shown here). However, even after the application of the calibration procedures, the relative error remains still relatively large. Moreover, the relative error obtained for case C.2 is large and may be considered as unacceptable; this suggests that one-year data may be inappropriate for forecasting purposes as regards the available offshore wind power potential. All these issues reveal the need for an in-depth assessment of the less reliable data sources in wind energy related applications. The same conclusion was obtained (in a different context) in Soukissian and Papadopoulos (2015a). See also the interesting discussion of Section 2.3 in Carta et al. (2013).

Table 4-6. Relative errors (%) of mean wind power density based on OLS, MM, M-H, LTS and L_1 for all the examined locations for cases C.1 and C.2.

Location	Method	Case C.1	Case C.2
Lesvos	OLS	38.37	40.74
	L_1	**26.71**	**29.25**
Mykonos	OLS	21.87	45.39
	LTS	**17.50**	**39.98**
Cabo Begur	OLS	28.54	57.84
	LTS	**23.45**	**42.41**
Cabo de Gata	OLS	25.84	13.21
	LTS	**16.91**	**6.66**

4.3.3   Calibration of directional data

Data and methodology used

Two Greek (Lesvos, Santorini) and two Spanish (Cabo Begur, Mahon) buoys are the reference data sources and the gridded data from the ERA-Interim (ERAI) and the Blended Sea Winds (BSW) datasets form the less accurate data sources (obtained from a NWP model and blending data from different satellites, respectively), that are calibrated. The closest grid point (belonging either to the model or the satellite product) to the coordinates of the examined buoy was chosen for the regression (and calibration) analysis. The data were firstly collocated in time with 6-hour common time frame (00:00, 06:00, 12:00, 18:00 UTC) for both combinations, i.e., buoy measurements and ERAI data, and buoy measurements and BSW data. The concurrent time series as regards the examined pairs buoy-ERAI and buoy-BSW are extending from 2004 to

2006 for the former pair, and from 2007 to 2009 for the latter one, since they provided the longest collocated triennial time span. The corresponding sample size along with the geographical coordinates for each buoy are presented in Table 4-7. Moreover, it was assumed that the change of wind direction due to the different reference heights is negligible.

Before proceeding with the application of the circular calibration on the available wind data, a basic statistical analysis for the three aforementioned data sources is provided. Wind directions are measured in degrees in the range $(0°, 360°]$ while the statistical and regression and calibration analysis is based on angles that are transformed in radians in the interval $(-\pi, \pi]$. Units are displayed, whenever necessary, if the values of the examined statistics are not in radians.

In Table 4-7, the main statistical parameters are summarized for each circular variable in the examined locations. It is noticed that the mean directions obtained from the concurrent wind data between buoy measurements and ERAI data are closer compared to the other pair of concurrent data (buoy and BSW) apart from Cabo Begur, where the corresponding difference is however rather low (~0.5°). The highest sample sizes are encountered in Santorini for both examined pairs of datasets. Moreover, the lowest values for V_θ and s_θ are found in Santorini and the highest values are depicted in Mahon. See also Figure 4-12 and Figure 4-13.

The prevailing wind directions obtained from the gridded wind data sources coincide in Lesvos and Santorini (NNE and N, respectively); see Figure 4-12. Furthermore, wind directions originating from the sector $[45°, 225°]$ are rather rare in the same locations. Regarding the Spanish locations in Figure 4-13, wind direction from all the examined data sources is much dispersed in Mahon; in Cabo Begur, the pattern of rose diagrams from the two examined pairs of concurrent wind directional data is relatively similar to each other compared to the other locations and the wind is almost unidirectional with prevailing direction coming from the NNW for the three wind data sources.

Table 4-7. Summary descriptive statistics from the concurrent wind directions (Buoy-ERAI, Buoy-BSW) for the Greek (Lesvos-LSV, Santorini-SNR) and Spanish (Cabo Begur-BGR, Mahon-MHN) locations.

Location	Data source	n	$\bar{\theta}$ (°)	$\bar{R}$ (−)	V_θ (−)	s_θ (−)
LSV	Buoy	3332	20.918	0.395	0.605	1.363
φ: 39.15°	ERAI		25.614	0.380	0.620	1.391
λ: 25.81°	Buoy	2972	21.365	0.390	0.610	1.372
	BSW		13.529	0.359	0.641	1.431
SNR	Buoy	4038	314.760	0.484	0.516	1.205
φ: 36.25°	ERAI		321.009	0.520	0.480	1.144
λ: 25.49°	Buoy	4028	314.855	0.484	0.516	1.205
	BSW		322.938	0.478	0.522	1.215
BGR	Buoy	3546	342.879	0.265	0.735	1.629
φ: 41.92°	ERAI		333.366	0.308	0.692	1.535
λ: 3.65°	Buoy	3494	343.140	0.263	0.737	1.635
	BSW		334.076	0.245	0.755	1.678
MHN	Buoy	3744	351.556	0.115	0.885	2.081
φ: 39.72°	ERAI		338.226	0.234	0.766	1.705
λ: 4.44°	Buoy	3694	353.139	0.113	0.887	2.089
	BSW		331.439	0.173	0.827	1.874

(a)

(b)
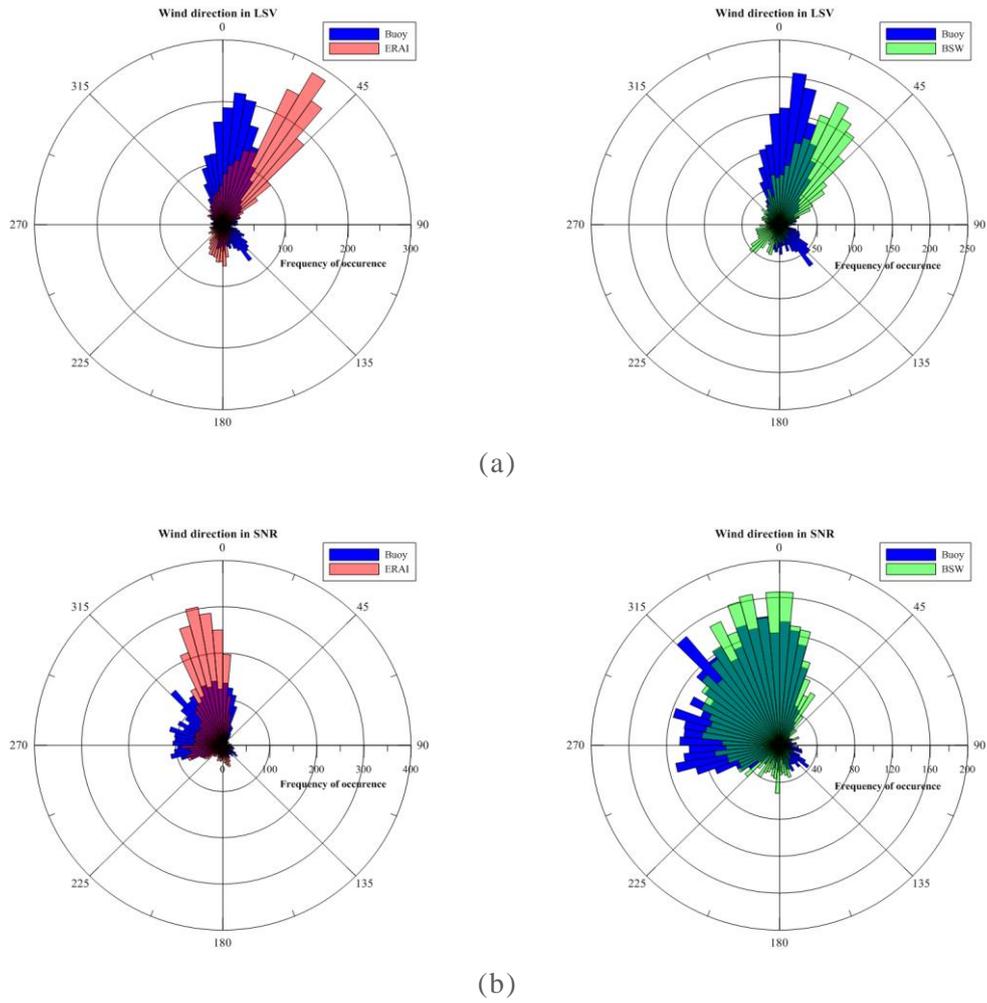
Figure 4-12. Rose diagrams of the two pairs of wind direction for(a) Lesvos and (b) Santorini.

(a)
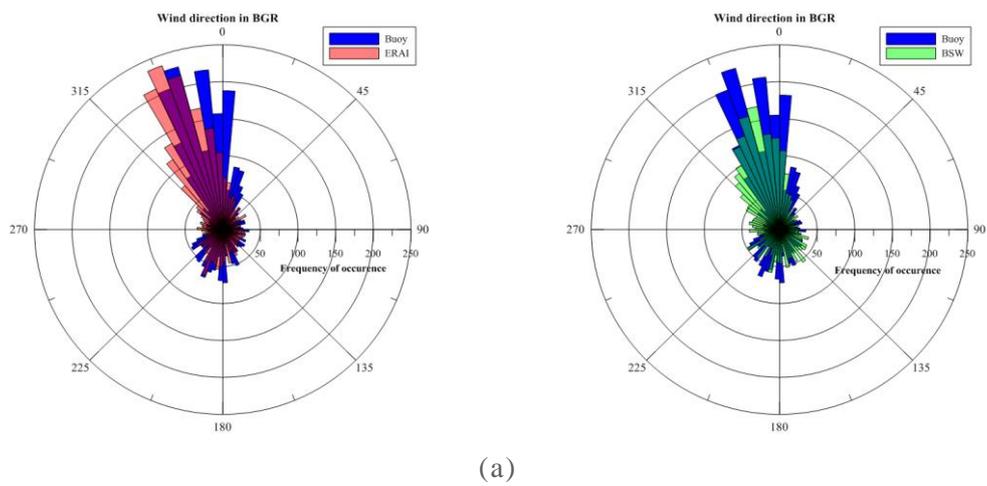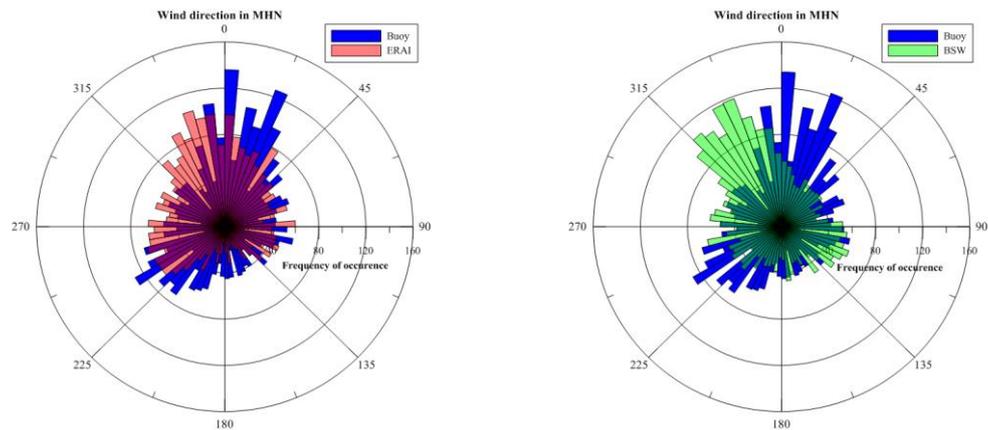
(b)

Figure 4-13. Rose diagrams of the two pairs of wind direction for (a) Cabo Begur and (b) Mahon.

After evaluating the performance of the examined regression/calibration methods presented in Section 2.6.2 with the statistical measures presented in Appendix B.4, the calibrated/corrected values of θ, denoted by $\hat{\theta}$, are used from the three applied calibration methods, i.e. classical circular calibration (AC1), inverse circular calibration (AC2), and circular calibration based on orthogonal distance (AC3).

In Table 4-8 and Table 4-9, the results obtained after the calculation of the abovementioned statistics for the different calibration techniques applied at each location are presented. The bold faced numbers denote the best value for each statistical criteria at the examined locations. The most striking result is that the inverse circular regression performs almost systematically better compared to the other two calibration methods. Specifically, the following results can be drawn:

1. As regards Table 4-8, from the overall combination of four locations and five statistical measures (i.e. 20 outcomes in total), AC2 performs better in 12 out of 20 outcomes, while AC1 and AC3 perform better 4 times each.
2. As regards Table 4-9, AC2 performs better in 11, AC1 in 5 and AC3 in 4 outcomes out of the total 20.
3. The values for MCAE and RME are systematically better for AC2, irrespectively of the gridded wind data source.
4. Overall, the lower values for BIAS, MCAE, RME and MRB are depicted after applying the calibration models for the ERAI results, apart from BIAS in Santorini and Mahon, and MRB in Lesvos.
5. All statistical criteria are fairly improved after applying the proposed circular calibration methods (apart from MRB in Lesvos), even if the data are not highly correlated/associated. This clearly suggests that circular calibration should be applied when an increased accuracy of wind direction is required.

Table 4-8. Statistical measures for the examined locations between buoy data and ERAI results before calibration (BC) and after calibration based on classical circular calibration (AC1), inverse circular calibration (AC2), and circular calibration based on orthogonal distance (AC3).

Location	Case	r_{cc}	BIAS	MCAE	RME	MRB
LSV	BC	0.4135	0.0820	0.6758	0.3090	−0.0454
	AC1	0.5230	**−0.0097**	0.6357	0.2974	0.0163
	AC2	**0.5336**	−0.0360	**0.5842**	**0.2848**	**0.0141**
	AC3	0.5259	−0.0450	0.5986	0.2880	0.0168
SNR	BC	0.7204	0.1091	0.4944	0.2279	−0.0776
	AC1	**0.7589**	0.0655	0.4611	0.2171	−0.0063
	AC2	0.7438	−0.0501	**0.4316**	**0.2078**	0.0101
	AC3	0.7521	**-0.0258**	0.4368	0.2100	**0.0017**
BGR	BC	0.6450	−0.1660	0.5259	0.2663	−0.0058
	AC1	0.6437	−0.0544	0.5346	0.2682	−0.0068
	AC2	**0.6478**	−0.0558	**0.5219**	**0.2653**	**−0.0015**
	AC3	0.6447	**−0.0346**	0.5257	0.2661	−0.0059
MHN	BC	0.8052	−0.2326	0.4527	0.2242	−0.0138
	AC1	0.8159	**−0.1149**	0.4477	0.2218	**0.0040**
	AC2	0.8169	−0.1653	**0.4409**	**0.2209**	0.0062
	AC3	**0.8171**	−0.1335	0.4431	0.2211	0.0048

Table 4-9. Statistical measures for the examined locations between buoy and BSW data before calibration (BC) and after calibration based on classical circular calibration (AC1), inverse circular calibration (AC2), and circular calibration based on orthogonal distance (AC3).

Location	Case	r_{cc}	BIAS	MCAE	RME	MRB
LSV	BC	0.0252	−0.1368	0.8193	0.3619	0.0003
	AC1	**0.4145**	0.1134	0.7559	0.3403	**0.0120**
	AC2	0.3529	−0.0685	**0.7098**	**0.3275**	0.0325
	AC3	0.3727	**−0.0291**	0.7220	0.3301	0.0276
SNR	BC	0.6610	0.1411	0.5999	0.2630	−0.0920
	AC1	**0.6850**	**−0.0144**	0.5357	0.2469	−0.0054
	AC2	0.6745	−0.0582	**0.4923**	**0.2330**	0.0174
	AC3	0.6835	−0.0512	0.5017	0.2367	**0.0050**
BGR	BC	0.4736	−0.1582	0.7256	0.3358	−0.0698
	AC1	0.5023	−0.1649	0.7217	0.3301	−0.0269
	AC2	**0.5193**	−0.1564	**0.6953**	**0.3225**	**−0.0055**
	AC3	0.5134	**−0.0802**	0.7025	0.3244	−0.0216
MHN	BC	0.6838	−0.3787	0.6107	0.2832	−0.0188
	AC1	0.7162	**−0.1071**	0.6092	0.2836	−0.0175
	AC2	0.7155	−0.2319	**0.6057**	**0.2825**	**−0.0116**
	AC3	**0.7179**	−0.1531	0.6065	0.2829	−0.0150

In the left panels of Figure 4-14, the situation before and after applying the calibration methods is presented for Lesvos location and both pairs of datasets by means of a scatter plot. This particular location was chosen in order to show the behaviour of AC2 method for two opposite situations: in the case of the pair "Buoy-ERAI", AC2 outperforms considerably the other methods (see also Table 4-8) while for the other pair of data the performance of AC2 is poorer (see also Table 4-9). With the reference line $y = x$ in mind, it is deduced that all calibration methods tend to shift more pairs towards this line with AC2 method reaching better results. A clearer conclusion is presented in the right panels of the same figure, where the histogram of the absolute circular distance is plotted for all the examined calibration methods; for clarity

purposes, only the central value of each bin (of 5° width) is depicted. The obtained results show that for smaller values of absolute circular distance, the most counts are provided for AC2 while as this value gets bigger (e.g. over 20°) then fewer counts correspond to this method. Finally, from these outcomes there is strong evidence that circular calibration should take part in studies where the accurate representation of direction is of importance.

Finally, in Figure 4-15 the calibrated wind directions for the best method is presented for each pair as regards Lesvos location. Comparing this figure with the corresponding ones before the calibration (see Figure 4-12(a)), it is once again verified that the calibrated values of wind direction obtained from the less accurate data sources are closer to the buoy measurements, even for the sectors with low frequency of occurrence.
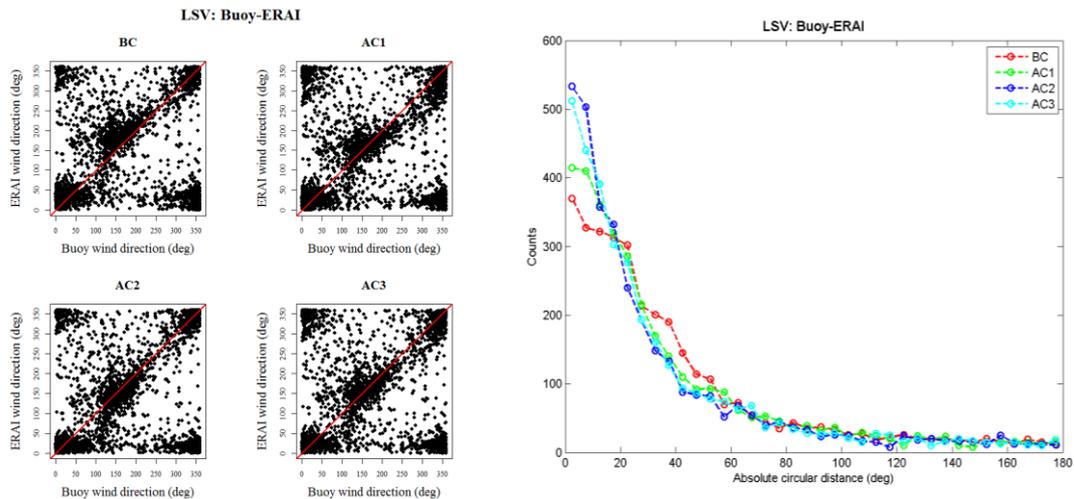
(a)
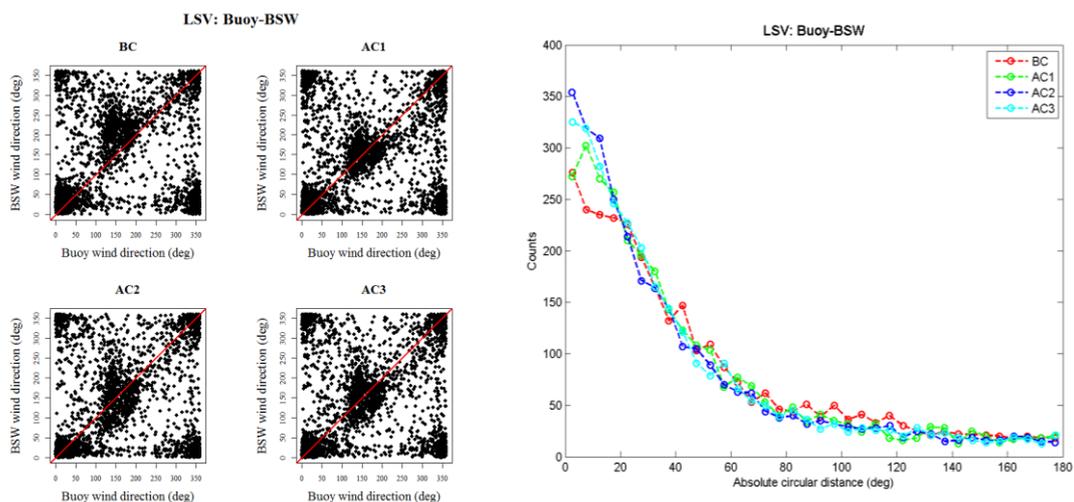
(b)

Figure 4-14. Calibration plots (left column) and histogram of absolute circular distance (right column) for (a) ERAI and (b) BSW data and all examined calibration methods in Lesvos.

4.3.4   Final comments

Linear data (e.g. wind speed data) from various sources quite often include erroneous observations that either can remain unnoticed or hidden in classical regression analysis or can be excluded from further assessment on the basis of some diagnostic tools. In either case, the results of the analysis may be highly misleading, since the presence of outliers or their false rejection can seriously affect the regression procedure (parameter estimation) and, consequently, the calibration results. Therefore, before proceeding to any analysis, the identification of outliers is an important, but rather delicate, procedure.
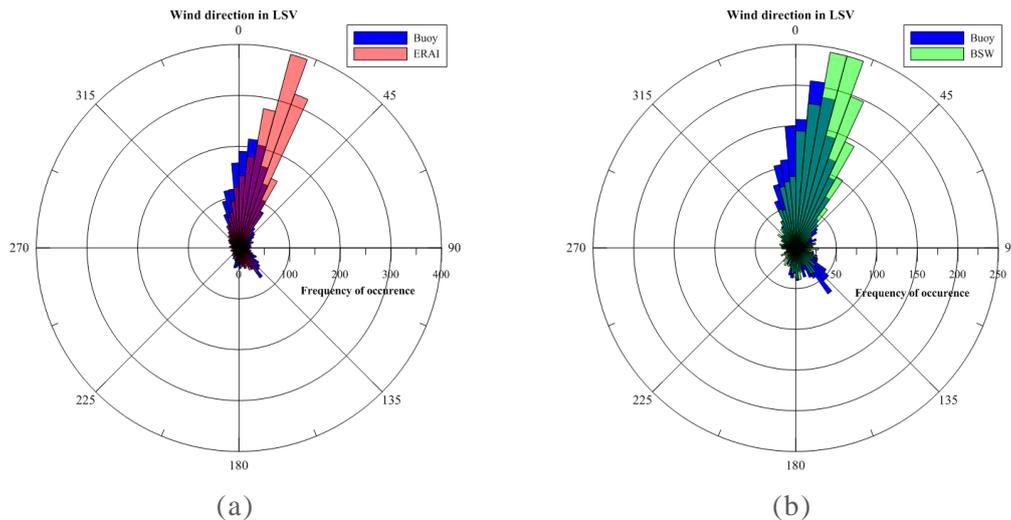
Figure 4-15. Rose diagrams after the calibration of wind direction with AC2 method of the pairs (a) buoy-model and (b) buoy-satellite for Lesvos.

In Section 4.3.2, robust estimators were described and applied for correcting wind speed measurements from less reliable data sources with reference to *in situ* measurements. Such measurements are prone to the presence of outliers and influential observations and as a consequence, the obtained results can be fallacious to derive decisions in wind speed assessment. Two different types of concurrent wind data sets referring to four offshore locations across the Mediterranean Sea were used: wind speed time series obtained from buoys and wind speed time series obtained from a high-resolution NWP model results covering various recording periods. The primary statistical analysis showed that wind speed is, in the mean, underestimated for the results of the examined atmospheric model compared to buoy measurements.

The evaluation of each robust method, along with the traditional OLS approach, was made by applying the regression and calibration procedure for different time periods (and consequently, different data samples). Using the entire available data sample (i.e. case C.1 above), the regression (and calibration) coefficients were estimated. In the other examined case (i.e. case C.2 above), an alternative methodology was applied for a more realistic and practical evaluation, where the regression (and calibration) relations were assessed from 'unused' wind measurements. Since there is not a unique statistical criterion for evaluating the performance of the examined calibration methods, several different statistical criteria were applied.

The obtained results from the evaluation procedures revealed that least trimmed squares (and secondarily, L_1-estimator) method performed systematically better for each examined case and for all locations than the rest methods. OLS method seemed to give rarely better results. Furthermore, the validity and performance of the regression/calibration methods was tested in the estimation of the mean wind power density. This assessment confirmed the results already

obtained from the evaluation of wind speed. Specifically, it was found that least trimmed squares and L_1-estimator methods performed systematically better than OLS, while the latter provided always the greatest relative estimation error.

Overall, robust statistics can provide a means for dealing efficiently with outliers in wind data samples in a theoretically justified way, while their implementation to wind energy assessment proved to give better results than classical OLS method. Least trimmed squares and L_1-estimator methods are characterized by their appealing definition and computability and can provide reasonable results, even if the outliers in the examined sample are numerous. Furthermore, another straightforward and efficient approach is to detect unusual wind speed data, with an emphasis on bad leverage points that have large residuals, through least trimmed squares analysis and then perform OLS regression without (all or a part of) these observations. Therefore, it is suggested that the use of robust methods should be seriously considered in wind energy related applications, since their effectiveness with samples containing outliers is indubitable.

In Section 4.3.3, three different circular regression/calibration models are proposed in order to correct wind data from less reliable wind data sources by using, as a reference source, measurements from oceanographic buoys. The examined data samples consist of two pairs of concurrent wind directions: 1) buoy measurements and results from the ERA-Interim data base, and; 2) buoy measurements and outputs from the Blended Sea Winds data base. Four locations, located in the Mediterranean Sea, are examined for each group with data covering a 3-year period.

The regression models were based on a tangent mapping while the parameter estimation was based on the minimization of circular distances. Moreover, it was assumed that the independent variable (i.e. wind direction from buoys) was error-free in contrast with the dependent one (i.e., wind direction from the gridded data sets). The evaluation of the proposed calibration models was based on five statistical criteria. The obtained results suggest that the inverse calibration, generally, performs better than the classical calibration and the calibration based on the orthogonal distance as regards the adopted statistical measures. In this respect, circular calibration should complement linear calibration, in cases where the accuracy is important such as wind energy assessment. Further work on the calibration of wind direction includes models that take into consideration errors in both variables or models that detect possible outliers.

4.4   Probabilistic modelling of metocean data

4.4.1   Synopsis

In this section, three families of models for the joint probabilistic description of wind speed and wind direction are examined and thoroughly evaluated, namely Johnson-Wehrly (JW) and two families of copulas (Farlie-Gumbel-Morgenstern (FGM) and Plackett (PLA)). These models are applied on long-term wind data obtained by two oceanographic buoys at different locations of the Mediterranean Sea, one in the Greek and the other one in the Spanish waters. The proposed bivariate models are theoretically sound and tractable, since they are defined by closed relations and are constructed by considering the marginal (univariate) distributions of wind speed and wind direction along with an appropriate dependence structure of the involved variables. In the univariate case, wind speed modelling is based on a wide range of conventional and mixture distributions, while wind direction is modelled through finite mixtures of von Mises distributions. The evaluation of the bivariate models is based on seven bin-specific goodness-of-fit criteria. The obtained results suggest that the performance of the JW model is rather superior, since it provides better fits compared to the other two families of bivariate distributions for the overwhelming majority of the examined cases and criteria. The most

efficient bivariate model is then implemented to estimate the detailed structure of wind power density at one selected location.

Part of the results presented in the following analysis have been published in:
Soukissian, T.H., Karathanasi, F.E., 2017. **On the selection of bivariate parametric models for wind data**. *Applied Energy* 188: 280–304.

4.4.2   Univariate and bivariate models for linear and directional variables

Univariate distributions for wind speed

Various univariate probability distributions can be used for fitting wind speed data; see Section 1.2.1. and references therein. In this analysis, the conventional parametric distributions, presented in Table 4-10, were initially evaluated using the Kolmogorov-Smirnov (K-S) and Anderson-Darling (A-D) goodness-of-fit tests.

After the pre-evaluation of these distributions (see Section 4.4.3 for more details), the most efficient distributions (i.e. those with an optimal performance regarding modelling of wind speed data in the examined locations) were identified and examined analytically, namely: Beta (BET), Burr (BUR), Dagum (DAG), Fatigue Life (FAL), Gamma (GAM), Generalized Extreme Value (GEV), Genelalized Gamma (GNG), Generalized Logistic (GNL), Generalized Pareto (GPA), Johnson SB (JSB), Kappa (KAP), Log-Logistic (LGL), Lognormal (LGN), Log-Pearson 3 (LP3), Pearson 6 (PE6), Pert (PER), Rayleigh (RAY), Wakeby (WAK) and Weibull (WEI) distributions. The definitions of the most usual probability distributions in wind energy assessment are presented in Section 1.2.1. Some of the distributions in the above table consider one or more location parameters. The consideration of such distributions is advocated by the fact that if these distributions provide a better fit than the location-free ones, then the former should be examined in the evaluation of the corresponding bivariate models.

For the better representation of wind regimes with particularities, apart from the conventional parameters, three parametric homogeneous and heterogeneous mixture distributions are additionally considered for wind speed modelling. Specifically, the homogeneous 2-parameter Weibull mixture (WW), the normal (truncated from below) mixture (NN), and the heterogeneous 2-parameter Weibull-Generalized Extreme Value mixture (WGEV) are examined (all with two components); for the corresponding definitions see also Section 1.2.1.

Table 4-10. Conventional parametric distributions for modelling wind speed.

Number of parameters	Conventional distributions
1	Rayleigh
2	Chi-Squared, Exponential, Gaussian, Levy, Log-Gamma, Nakagami, Pareto, Rayleigh, Reciprocal, Rice, Uniform
3	Erlang, Fatigue Life, Fréchet, Gamma, Generalized Extreme Value, Generalized Logistic, Generalized Pareto, Inverse Gaussian, Log-Logistic, Lognormal, Log-Pearson 3, Pearson 5, Pert, Power function, Weibull
4	Beta, Burr, Dagum, Generalized Gamma, Johnson S$_B$, Kappa, Kumaraswamy, Pearson 6
5	Wakeby

<u>Univariate distributions for wind speed</u>

As regards the directional variable (i.e. wind direction), a finite mixture of von Mises (vM) distributions is implemented; see Section 1.2.2 for the definition. In respect with the problem of selecting the number of components for describing this finite mixture model in an easy and fast way, it is still an open issue. In this study, the Bayesian information criterion (BIC), proposed by Schwarz (1978), was applied in order to select the optimal number of components for each finite mixture model. BIC is defined by

$$\text{BIC} = -2 \log L + b \ln N, \tag{4.7}$$

where L is the maximized value of the likelihood function for the mixture model, and b is the number of parameters in the mixture model. At the right-hand side of Eq. (4.7), the first term is a measure of lack of fit and the second one measures the degree of complexity of the model. Moreover, the algorithm proposed by Garcia-Portugues (2013) was applied for selecting the range of the number of components; the number of components that minimized the BIC function was selected as the most optimal value for the estimation of the parameters.

In regard of the bivariate case of modelling wind speed and wind direction, the Johnson-Wehrly (JW), the Farlie-Gumbel-Morgenstern (FGM) and the Plackett families of distributions are assessed; for a brief theoretical background on these distributions see Section 1.3.1.

4.4.3 Goodness-of-fit testing

Regarding wind speed, a preliminary selection of the most efficient distributions, out of the 36 conventional distributions that were initially examined (see Table 4-10), was based on the Kolmogorov–Smirnov (K-S) and Anderson–Darling (A-D) goodness-of-fit tests. K-S test is non-parametric and is based on the absolute deviations (largest vertical distances) between the empirical distribution function (i.e. sample cdf) and the specified hypothetical continuous cdf. The main disadvantage of this test is its sensitivity near the centre of the distribution and that the distribution must be specified. A modification of K-S test is A-D test that is used to verify if the sample data comes from a population with a specific distribution. The critical values of A-D test are dependent on the specified distribution that is being tested in contrast to K-S test, allowing in this way a more sensitive test. A-D test gives more weight to deviations at the tails of distributions than K-S test. The two tests are not equivalent and may generate inconsistent indications of fit performance among the candidate pdfs; see, e.g. Chang (2011), and Soukissian (2013). Therefore, the ten best-fit distributions, characterized by the smallest test statistics, according to K-S test and the corresponding ones according to A-D tests were identified for the examined locations. Nevertheless, the final number of the analytically examined distributions for each location was less than 20 due to several overlaps between the best-fit distributions provided by the two tests; see also Section 4.4.4.

In order to characterize the obtained fits in a uniform way, the final evaluation of the distributions was based on the coefficient of determination $R_{a,1}^2$ (the lower index '1' denotes the univariate case), which quantifies the association between the observed cumulative probabilities and the predicted cumulative probabilities of a wind speed distribution. For each definition, see Appendix B.5. This coefficient has been adopted in many wind speed modelling studies; see, for example, Carta et al. (2009) and references therein, Ouarda et al. (2015).

On the other hand, the evaluation of the bivariate (and multivariate, in general) models is not a trivial issue, since the available statistical tools are rather poor (McAssey, 2013). Although the literature for multi-normal evaluation is abundant, the corresponding tests cannot be directly extended in the general multivariate case, not even for the bivariate one. McAssey (2013) also

states that some attempts for establishing goodness-of-fit tests in multiple dimensions are either extremely difficult to compute or intractable for most multivariate distributions. In the present analysis, the evaluation of the obtained bivariate fits was based on seven bin-specific measures of performance, which are presented here due to their rare appearance in the corresponding literature.

Let $p_{ij}^{(o)}$ denote the fraction of points (from the real data set) that belong in the (i,j) −bin (cell) and $p_{ij}^{(e)}$ the corresponding fraction of points from the estimated bivariate distribution. Let also $N_{ij}^{(o)}$ denote the observed number of data points falling in the (i,j) −cell. A data point (x,θ) belongs in the (i,j) −cell if $(x_i < x \le x_{i+1})$ and $(\theta_j < \theta \le \theta_{j+1})$ where x_i, θ_j, $i = 1,2,\dots,I$, $j = 1,2,\dots,m$, are the corresponding cell boundaries.

For the estimated (theoretical) bivariate pdf, the probability that an observation falls in this cell is evaluated as follows:

$$Pr[(x,\theta) \in cell(i,j)] = p_{ij}^{(e)} = F_{X,\Theta}(x_{i+1}, \theta_{j+1}) - F_{X,\Theta}(x_{i+1}, \theta_j)$$
$$- F_{X,\Theta}(x_i, \theta_{j+1}) + F_{X,\Theta}(x_i, \theta_j). \tag{4.8}$$

The root mean square error (RMSE) is defined as

$$RMSE = \sqrt{\frac{SS_E}{N_T}}, \tag{4.9}$$

where $SS_E = \sum_{i,j}\left(p_{ij}^{(o)} - p_{ij}^{(e)}\right)^2$ is the sum of squared error, which measures the total difference between the observed and the expected frequency for all bins, and N_T is the total number of bins.

The relative root mean square error (RRMSE) is defined as

$$RRMSE = \sqrt{\frac{SRS_E}{N_T}}, \tag{4.10}$$

where $SRS_E = \sum_{i,j}\left(\frac{p_{ij}^{(o)} - p_{ij}^{(e)}}{p_{ij}^{(o)}}\right)^2 = \sum_{i,j}\left(1 - p_{ij}^{(e)}\right)^2$ is the sum of relative squared error.

The mean absolute error (MAE) is defined as

$$MAE = \frac{1}{N_T}\sum_{i,j}\left|p_{ij}^{(o)} - p_{ij}^{(e)}\right|. \tag{4.11}$$

The index of agreement (IA), suggested by Zhou et al. (2010), is defined as

$$IA = 1 - \frac{\sum_{i,j}\left(p_{ij}^{(o)} - p_{ij}^{(e)}\right)^2}{\sum_{i,j}\left(\left|p_{ij}'^{(o)}\right| - \left|p_{ij}'^{(e)}\right|\right)^2}, IA \in [0,1], \tag{4.12}$$

where $p_{ij}'^{(o)}$, $p_{ij}'^{(e)}$ denotes the difference between the observed and the mean frequency, and the difference between the estimated and the mean frequency, respectively. Willmott (1982) introduced this statistic while Webb et al. (2009) applied this index to wind erosion analysis.

The χ^2 statistic (also called χ^2 −error) is defined as

$$\chi^2 = \sum_{i,j} \frac{\left(np_{ij}^{(o)} - np_{ij}^{(e)}\right)^2}{np_{ij}^{(e)}} = \sum_{i,j} \frac{\left(N_{ij}^{(o)} - np_{ij}^{(e)}\right)^2}{np_{ij}^{(e)}}, \tag{4.13}$$

where n denotes the total number of observations.

The adjusted coefficient of determination $R_{a,2}^2$ (lower index '2' denotes the bivariate case) measures the strength of the linear relationship between the expected and the observed frequencies of all bins and is defined as

$$R_{a,2}^2 = 1 - \frac{(N-1)\text{SS}_\text{E}}{(N-q-1)\text{SS}_\text{T}}, \tag{4.14}$$

where q is the number of parameters estimated for the particular distribution, N is the total number of non-empty bins and SS_T denotes the total sum of squares, reflecting the total difference between the observed frequencies and the average frequency $\overline{p}^{(o)}$ for all bins, i.e. $\text{SS}_\text{T} = \sum_{i,j} \left(p_{ij}^{(o)} - \overline{p}^{(o)}\right)^2$.

The bivariate pdfs providing the smallest values of RMSE, MAE, RRMSE and χ^2, and the highest values of $R_{a,2}^2$ and IA are likely to be characterized by a smaller fit error.

Moreover, Mathisen and Bitner-Gregersen (1990) proposed a more sophisticated method for the evaluation of the bivariate fits based on the normalized deviation d_{ij} between the observed number of data points $N_{ij}^{(o)}$ falling in a cell and the expected number of points $np_{ij}^{(e)}$. d_{ij} is provided by

$$d_{ij} = \frac{N_{ij}^{(o)} - np_{ij}^{(e)}}{\sigma_{ij}^{(e)}}, \tag{4.15}$$

where $\sigma_{ij}^{(e)}$ is the normalizing factor. $\sigma_{ij}^{(e)}$ is the expected standard deviation for the number of data points falling in the cell according to the binomial distribution, since each outcome may be considered as 'success' (i.e. the data point lies inside the examined cell) or 'failure'; therefore, $\sigma_{ij}^{(e)} = \sqrt{np_{ij}^{(e)}\left(1 - p_{ij}^{(e)}\right)}$. Values of d_{ij} close to zero indicate a good fit of the adopted model. In addition, positive values denote that the fitted model assigns a lower probability compared to the actual data in the specific cell; the opposite holds true for negative values.

The main conceptual difference between the different fit performance measures used here is that d_{ij} provides a characterization of the performance of the estimated analytic bivariate pdf in the entire domain of (x, θ) while the other measures provide a unique characterization for the performance of the model. Thus, d_{ij} identifies the particular areas in the (x, θ) −plane where the bivariate model either performs well or underperforms.

4.4.4   Data and numerical results

Long-term measured wind speed and direction data from one Greek oceanographic buoy of the POSEIDON marine monitoring network and one Spanish oceanographic buoys of the Spanish Port Authority (Puertos del Estado) are presented in this analysis; their exact location and measurement period are shown in Table 4-11. Three more buoys and one onshore meteorological mast for wind measurements of the Centre for Renewable Energy Sources and Saving (CRES) were additionally included in Soukissian and Karathanasi (2017). Before the statistical analysis and the parameter estimation procedure, the wind data were corrected and filtered, negative and stacked values were removed, while only concurrent measurements of wind speed and wind direction were taken into consideration.

Table 4-11. Station names, geographical locations, recording periods and sample sizes of wind data sets.

Name location	Latitude, longitude (°)	Measurement period	n
Crete	35.79°N, 24.92°E	2007–2015	12,665
Mar de Alboran	36.27°N, 5.03°W	1997–2006	29,100

In Table 4-12 and Table 4-13, the main statistical parameters of wind speed u (e.g. standard deviation s_u, coefficient of variation CV_u, skewness Sk_u, kurtosis Ku_u) and wind direction θ from the above data sets are presented, respectively. The most intense wind climate corresponds to Crete, with mean wind speed m_u =5.414 m/s and overall maximum value of wind speed $\max_u$ =18.75 m/s.

Table 4-12. Basic (linear) statistics of wind speed at the examined locations.

Name location	m_u (m/s)	$\min_u$ (m/s)	$\max_u$ (m/s)	s_u (m/s)	CV_u (%)	Sk_u (–)	Ku_u (–)
Crete	5.414	0.195	18.750	2.809	51.874	0.567	0.321
Mar de Alboran	5.273	0.200	16.900	2.949	55.926	0.368	−0.553

As regards the corresponding results for wind direction, it is noticed that the winds blow, in the mean, from the WNW sector for both examined locations. The highest value of mean resultant length $\overline{R}_\theta$ is depicted at Crete, suggesting rather concentrated data while highest values of circular variance V_θ (and s_θ) corresponds to Mar de Alboran denoting that wind direction in this location is more uniformly distributed on the circle compared with the Greek location. Finally, the highest value, in the absolute sense, of circular skewness Sk_θ is encountered at Mar de Alboran (−0.493) denoting that the corresponding dataset is rather non-unimodal.

Table 4-13. Basic (circular) statistics of wind direction at the examined locations.

Name location	m_θ (rad)	$\overline{R}_\theta$ (–)	V_θ (–)	s_θ (–)	Sk_θ (–)	Ku_θ (–)
Crete	5.285	0.396	0.604	1.099	−0.041	0.269
Mar de Alboran	5.059	0.097	0.903	1.344	−0.493	0.207

Univariate distributions for wind speed

For each examined location the preliminary selection of the distributions that are further analysed was made by using K-S and A-D goodness-of-fit tests as mentioned in Section 4.4.3. These distributions along with the corresponding K-S and A-D test statistics are summarized in Table 4-14. In the same table, the corresponding values for the mixture distributions are also

provided (i.e. WW), WGEV and NN). Empty cells denote that the provided test statistic values are too large.

Overall, 19 conventional distributions and three mixture distributions with two components were further evaluated by means of $R_{a,1}^2$. The corresponding results, shown in Figure 4-16, are plotted appropriately scaled, i.e. using $\left|\log\left(1 - R_{a,1}^2\right)\right|$ instead of $R_{a,1}^2$, since the values of the latter are often very close to each other, rendering their differences imperceptible in a graph.

The results of the evaluation based on $R_{a,1}^2$ suggest the following: as regards the conventional distributions, WAK distribution provided the best fit for both locations, KAP distribution provided the second best fit again for both locations while GEV and GNG provided the third best fit for Crete and Mar de Alboran, respectively. As regards the mixture distributions, NN provided a better fit than WGEV and WW for Crete and WW provided a better fit than the other two for Mar de Alboran.

Table 4-14. Kolmogorov-Smirnov and Anderson-Darling test statistics.

Distributions	Test	Crete	Mar de Alboran	Distributions	Test	Crete	Mar de Alboran
BET	K-S	0.032	0.041	LGL	K-S	0.027	
	A-D	11.026	62.371		A-D	14.965	
BUR	K-S	0.03	0.050	LGN	K-S	0.026	
	A-D	9.807	117.03		A-D	7.058	
DAG	K-S	0.030	0.033	LP3	K-S		0.041
	A-D	9.963	35.171		A-D		77.784
FAL	K-S	0.027		NN	K-S	0.030	0.030
	A-D	7.279			A-D	2.495	7.852
GAM	K-S	0.029		PE6	K-S	0.026	
	A-D	7.930			A-D	7.639	
GEV	K-S	0.023	0.049	PER	K-S		0.044
	A-D	5.689	118.88		A-D		69.657
GNG	K-S		0.026	RAY	K-S		0.051
	A-D		20.151		A-D		129.05
GNL	K-S	0.029		WAK	K-S	0.016	0.013
	A-D	17.36			A-D	419.47	1504.9
GPA	K-S		0.030	WEI	K-S	0.032	0.056
	A-D		2933.1		A-D	10.906	120.88
JSB	K-S	0.025	0.029	WGEV	K-S	0.030	0.024
	A-D	6.318	126.24		A-D	3.340	4.823
KAP	K-S	0.027	0.020	WW	K-S		0.021
	A-D	5.676	25.036		A-D		4.149

(a)   (b)
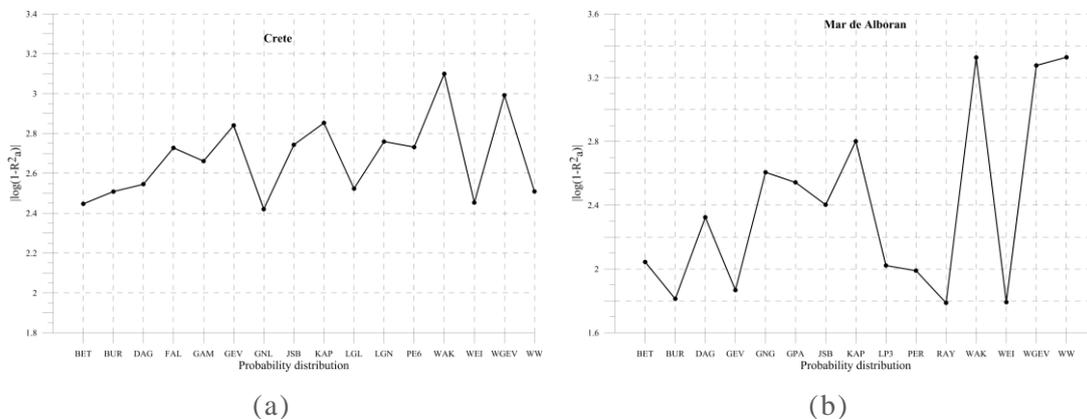
Figure 4-16. Values of coefficient of determination $R_{a,1}^2$ for the corresponding probability distribution functions for (a) Crete and (b) Mar de Alboran.

Summing up, the six best fits for each location that are analytically assessed in the bivariate case, are summarized as follows (numbers in parentheses denote the corresponding values of $R_{a,1}^2$):

- Crete: NN (0.99961), WAK (0.99960), WGEV (0.9995), KAP (0.99930), GEV (0.99928), LGN (0.9991);
- Mar de Alboran: WW (0.9998), WAK (0.99976), WGEV (0.9997), KAP (0.9992), NN (0.99883), GNG (0.99880).

Notice that KAP and WAK distributions appear between the best-fit distributions for the two examined locations. In Figure 4-17, the histograms of wind speed along with the six best fit pdfs are plotted for the examined locations.

Univariate distributions for wind direction

In Table 4-15, the parameters of the mixtures of vM distributions (see Eq. (1.36)) for wind direction modelling are summarized, providing also the weighting parameter. Based on the BIC criterion, Crete and Mar de Alboran were described with four and three components, respectively.
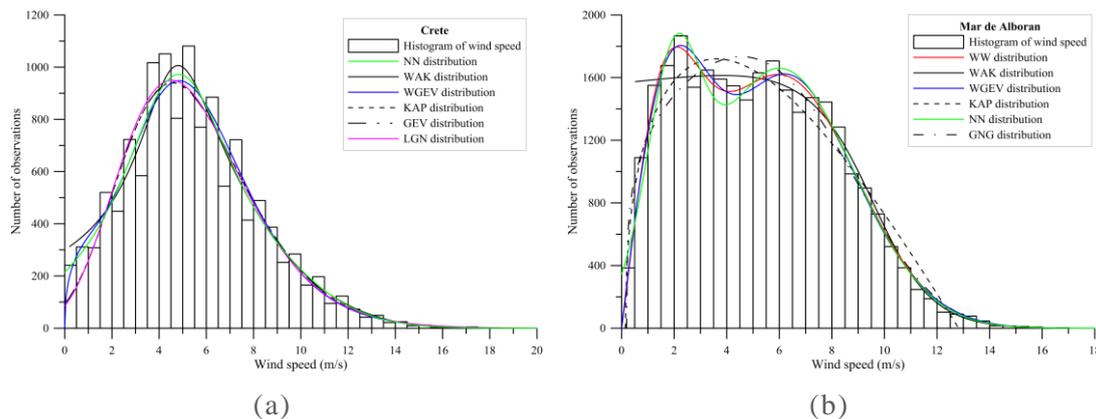
Figure 4-17. Histograms of wind speed along with the best fits from the corresponding probability density function for (a) Crete and (b) Mar de Alboran.

Table 4-15. Parameters of the best-fit wind direction distributions used in bivariate modelling. κ and ω parameters are dimensionless and μ parameter is in rad.

Parameters	Crete	Mar de Alboran
$\kappa_i, i-1,2,...$	0.499	6.517
	1.359	5.361
	5.778	0.406
	8.964	
$\mu_i, i-1,2,...$	2.665	1.236
	2.007	4.544
	5.918	1.426
	4.824	
$\omega_i, i-1,2,...$	0.176	0.321
	0.164	0.464
	0.298	0.215
	0.362	

In Figure 4-18, the histograms of wind direction along with the vM mixture distributions are plotted for the examined locations. From the shape of these histograms, it is evident that the use of mixture of circular vM distributions for modelling wind direction at the examined locations is inevitable.
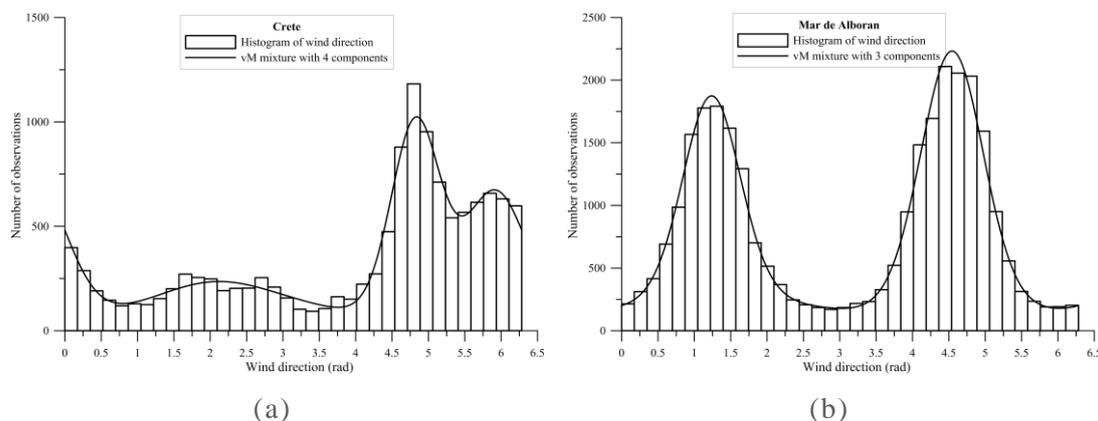
Figure 4-18. Histograms of wind direction along with the fitted mixtures of vM distributions for (a) Crete and (b) Mar de Alboran.

Bivariate distributions of wind speed and direction

The values of the parameters r_{FGM} and ψ_P involved in the FGM (see Eq. (1.43)) and PLA (see Eq. (1.48)) family of distributions, respectively, are shown in Table 4-16. From the obtained values of r_{FGM} it is concluded that the FGM model can be safely applied for the two locations, since $-1/3 \leq r_{\text{FGM}} \leq 1/3$.

Table 4-16. Parameters of the bivariate models FGM and PLA.

Parameters	Crete	Mar de Alboran
r_{FGM}	0.106	0.263
ψ_P	1.150	1.208

Regarding the JW family, in Table 4-17, the parameters of the vM mixture for modelling ψ variable, see Eq. (1.40), are provided for the examined locations and pdfs of wind speed.

Evaluation of the fitted bivariate distributions

For the evaluation of the obtained bivariate distributions RMSE, RRMSE, MAE, IA, χ^2 and $R^2_{a,2}$, along with the deviance statistic, are calculated. In Table 4-18 and Table 4-19, the values of the above measures are shown for the six best fit wind speed pdfs for the examined locations. Since the involved quantities in the estimation of these measures are expressed through frequencies (χ^2 error represents 'counts'), all presented measures are dimensionless. The wind speed distributions in the tables are sorted in a decreasing order according to the values of the $R^2_{a,1}$ criterion. Boldface numbers denote the best value of the particular measure for each bivariate family, while both italics and boldface numbers denote the overall best value of the particular measure for all bivariate families. Furthermore, in Figure 4-19, the deviance statistic is plotted for each bivariate family.

Table 4-17. Parameters of the vM mixture of distributions for ψ variable of the JW family for the examined locations. κ_1, κ_2, ω_1, ω_2 parameters are dimensionless and μ_1, μ_2 parameters are in rad.

Location		**Probability distributions and parameters**					
		GEV	KAP	LGN	NN	WAK	WGEV
	κ_1, κ_2	0.758	0.466	0.768	0.464	0.782	0.802
		0.372	0.747	0.464	0.769	0.478	0.404
Crete	μ_1, μ_2	5.709	2.938	5.710	2.972	5.755	5.768
		3.001	5.704	2.939	5.734	2.982	3.072
	ω_1, ω_2	0.503	0.453	0.539	0.467	0.530	0.491
		0.497	0.547	0.461	0.533	0.470	0.509
		GNG	KAP	NN	WAK	WGEV	WW
	κ_1, κ_2	0.348	0.218	0.762	0.229	0.746	0.274
		0.844	0.772	0.338	0.762	0.335	0.791
Mar de Alboran	μ_1, μ_2	3.824	3.658	0.686	4.082	0.662	3.806
		0.642	0.583	3.875	0.735	3.812	0.658
	ω_1, ω_2	0.467	0.494	0.562	0.487	0.567	0.478
		0.5333	0.506	0.438	0.513	0.433	0.522

Some conclusions as regards the general behaviour of the three examined bivariate approaches can be summarized as follows:

i. The overall largest values of $R_{a,2}^2$ and IA, and the overall smallest values of RMSE, MAE, RRMSE (except for Mar de Alboran), and χ^2 statistic (except for Mar de Alboran) are consistently provided by the JW family, i.e. JW family provides the optimal values for the test statistics in 16 out of 18 cases. Although the differences between the values of these measures with respect to the examined bivariate families are not large, JW family provides consistently the optimal values. This result suggest the superiority of the JW family in comparison to the FGM and the PLA families of distributions.

ii. FGM family performs better than PLA family with respect to all measures of evaluation for Mar de Alboran while PLA family performs better than FGM family with respect RMSE, MAE, χ^2 and $R_{a,2}^2$ for Crete.

iii. For Crete, the best-fitting models from all the examined families include mixture distributions for wind speed.
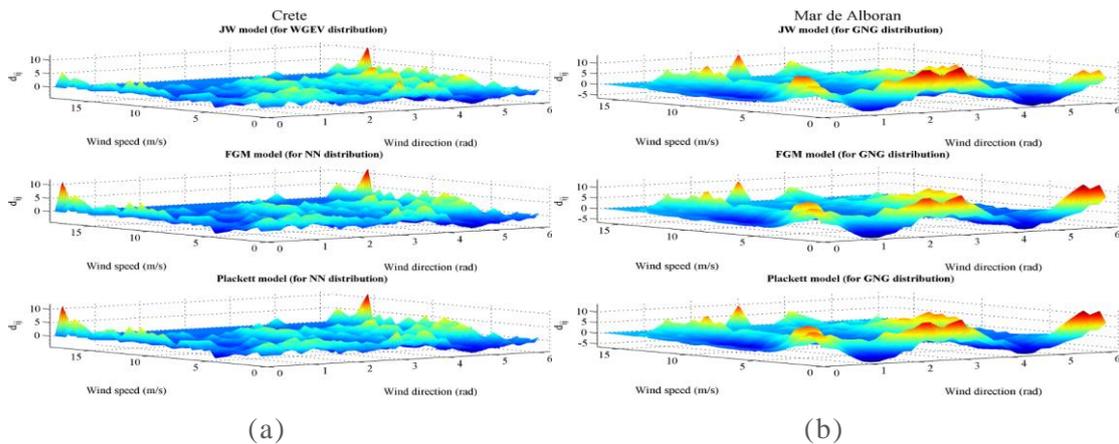
Figure 4-19. Normalized deviations for JW, FGM and PLA bivariate families for (a) Crete and (b) Mar de Alboran.

Table 4-18. Values of RMSE, RRMSE, MAE, IA, χ^2 and $R^2_{a,2}$ for JW, FGM and PLA families of bivariate distributions for Crete. All measures are dimensionless.

u pdf	JW						FGM						PLA					
	RMSE	RRMSE	MAE	IA	χ^2	$R^2_{a,2}$	RMSE	RRMSE	MAE	IA	χ^2	$R^2_{a,2}$	RMSE	RRMSE	MAE	IA	χ^2	$R^2_{a,2}$
NN	6.021	1.007	3.375	0.953	2177	0.834	6.246	1.113	3.508	0.947	2316	0.822	6.227	1.118	3.493	0.947	2313	0.823
WAK	6.103	1.007	3.400	0.951	2243	0.830	6.298	1.117	3.536	0.946	2376	0.819	6.277	1.122	3.519	0.946	2369	0.820
WGEV	6.014	1.005	3.392	0.953	2177	0.835	6.265	1.110	3.521	0.946	2324	0.820	6.245	1.115	3.505	0.947	2309	0.822
KAP	6.089	1.009	3.432	0.952	2179	0.831	6.344	1.116	3.571	0.945	2310	0.816	6.324	1.121	3.555	0.946	2288	0.817
GEV	6.092	1.013	3.432	0.952	2177	0.830	6.354	1.119	3.572	0.945	2308	0.815	6.334	1.125	3.556	0.945	2285	0.816
LGN	6.127	0.985	3.450	0.951	2157	0.828	6.365	1.091	3.592	0.945	2320	0.815	6.343	1.096	3.576	0.946	2306	0.816

Table 4-19. Values of RMSE, RRMSE, MAE, IA, χ^2 and $R^2_{a,2}$ for JW, FGM and PLA families of bivariate distributions for Mar de Alboran. All measures are dimensionless.

u pdf	JW						FGM						PLA					
	RMSE	RRMSE	MAE	IA	χ^2	$R^2_{a,2}$	RMSE	RRMSE	MAE	IA	χ^2	$R^2_{a,2}$	RMSE	RRMSE	MAE	IA	χ^2	$R^2_{a,2}$
WW	15.356	1.669	9.030	0.944	7688	0.807	15.978	1.584	9.263	0.938	7772	0.791	16.390	1.642	9.478	0.934	7975	0.780
WAK	15.451	1.718	9.064	0.943	7377	0.805	16.013	1.612	9.350	0.937	7366	0.790	16.419	1.674	9.567	0.934	7578	0.779
WGEV	15.387	1.645	9.056	0.943	7795	0.806	15.975	1.570	9.230	0.938	7862	0.791	16.389	1.626	9.447	0.934	8103	0.780
KAP	15.467	1.820	9.292	0.942	8182	0.804	16.156	1.720	9.565	0.936	8197	0.786	16.560	1.790	9.783	0.932	8371	0.775
NN	15.391	1.657	9.040	0.943	8012	0.806	16.004	1.581	9.230	0.938	8006	0.790	16.418	1.638	9.442	0.934	8331	0.779
GNG	15.218	1.662	9.215	0.945	8012	0.811	15.842	1.591	9.502	0.939	8374	0.795	16.256	1.649	9.727	0.935	9489	0.784

As regards the specific performance of the JW, FGM and PLA families with respect to the univariate distributions of wind speed, the following conclusions can be drawn:

i. The statistical measures examined do not provide, in general, compatible suggestions of the bivariate distribution fit performance. The issue of inconsistency between different statistical measures is known for the univariate evaluation of goodness of fit; see, for example, the relevant discussion in Zhou et al. (2010). Thus, the selection of the best fit is confined to the number (multitude) of criteria suggesting the optimal fit.

ii. In general, the bivariate best-fit distribution is not provided by the corresponding univariate best-fit distribution for wind speed.

iii. For Crete, the bivariate best-fit distribution is provided by the JW family for the WGEV mixture (RMSE, MAE, IA and $R_{a,2}^2$ provide the overall optimal values).

iv. For Mar de Alboran, the bivariate best-fit distribution is provided by the JW family for the GNG distribution (RMSE, IA and $R_{a,2}^2$ provide the overall optimal values).

In order to evaluate in further detail the comparative performance of JW, FGM and PLA families in the entire (u, θ) −plane, the normalized deviation d_{ij} is presented in Figure 4-19 for the best bivariate fit for each family, according to the selection presented above. The values of the normalized deviation that are close to zero indicate a good fit of the adopted model. Negative and positive values indicate overestimation and underestimation, respectively, of the probability mass of the particular cell by the fitted distribution. According to the examined bivariate models and locations, the overall range of d_{ij} is between −5 and 12. Specifically, for Crete, all models underperform for wind directions around the northern sector with simultaneous high values of wind speed while for Mar de Alboran, it seems that the bivariate models underperform (although with a variable intensity) in an extended area of (u, θ) combinations. Overall, the poor performance of the bivariate models is more evident for the FGM and PLA families. The obtained results suggest that JW model performs better compared to the other two approaches; although the pattern of d_{ij} provided by the three families is almost identical for all the examined locations, the corresponding values are lower for the JW model. In conclusion, taking into consideration the results provided by the seven metrics used in this analysis, the JW family seems to provide consistently the best fits and may be used as a solid base for bivariate modelling of wind speed and direction (compared to FGM and PLA families). A potential improvement in the results of the JW model may be based on the consideration of non-negative trigonometric sums for the modelling of $f_\Psi(\psi)$. As noted by Fernández-Durán (2007), the use of this representation may provide flexibility in the modelling of different (even of very complex) dependence structures.

In Figure 4-20–Figure 4-22, the selected bivariate best-fit pdfs of wind speed and wind direction according to JW, FGM and PLA families of distributions for the examined locations are presented. In general, the patterns of the FGM and PLA models have a strong resemblance. As regards wind direction, the location of the peaks from the univariate vM mixture model is very close to the corresponding one of the bivariate case at each examined site. Similarly, as regards wind speed, the highest values of $f_{U,\Theta}$ are depicted in the ranges that are consistent with the histograms of Figure 4-17. Moreover, FGM and PLA models exhibit, in general, smoother shapes for $f_{U,\Theta}$ than JW family at each location.
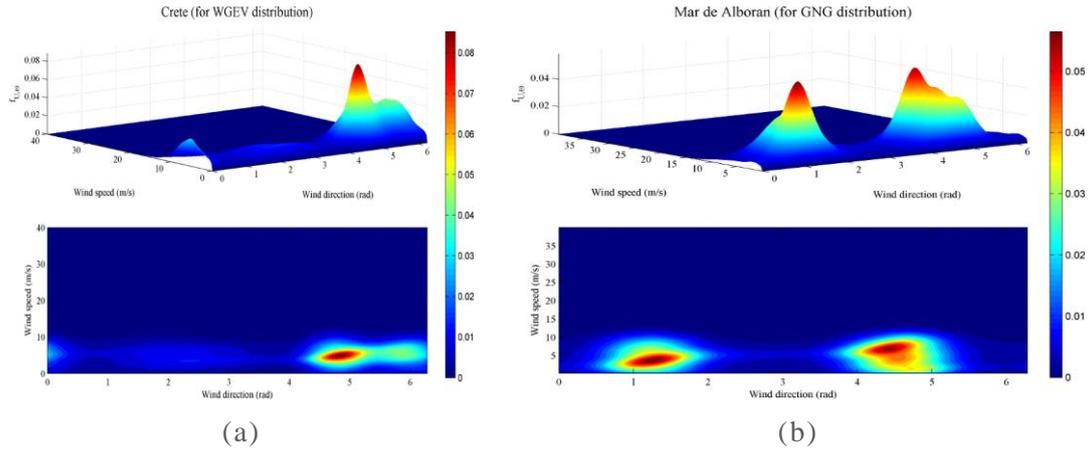
Figure 4-20. Best-fit for bivariate pdfs of wind speed and wind direction for JW family for (a) Crete and (b) Mar de Alboran.
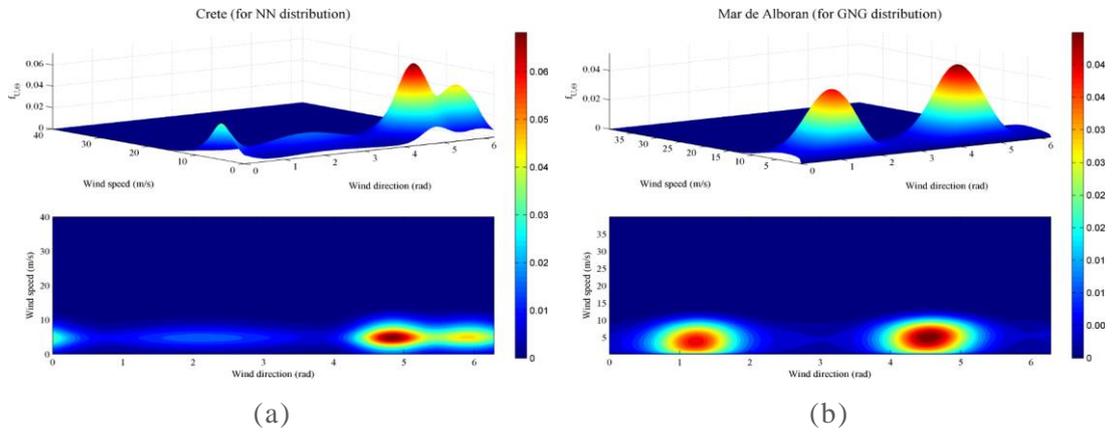
Figure 4-21. Best-fit for bivariate pdfs of wind speed and wind direction for FGM family for (a) Crete and (b) Mar de Alboran.
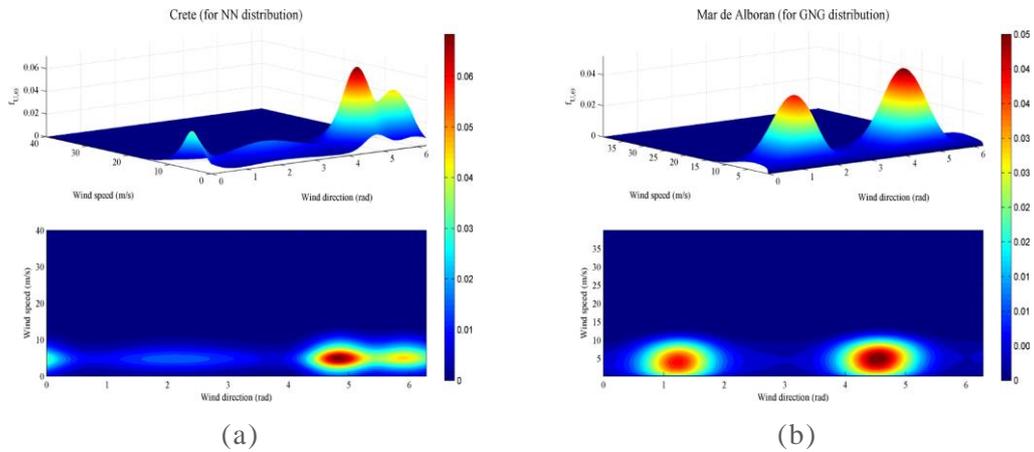
Figure 4-22. Best-fit for bivariate pdfs of wind speed and wind direction for PLA family for (a) Crete and (b) Mar de Alboran.

4.4.5   Application in wind energy assessment

The preliminary assessment of the available offshore wind power potential in an area requires the estimation of the theoretical wind power density per rotor swept area. For a homogeneous, isotropic, free-stream air flow with wind speed $u = u_\infty$, the theoretical wind power density P_0 is given by

$$P_0 = \frac{1}{2}\rho u_\infty^3. \tag{4.16}$$

Due to the temporal and spatial variability of wind speed, wind direction and air density, more detailed information as regards wind power for different occurrences of these variables at the specific area is required during the stage of wind farm design and layout. In this respect, for the evaluation of wind energy at a particular location, the wind power density distribution is used. This quantity is estimated by

$$e(u, \theta, \rho) = \frac{1}{2}\rho u^3 f_{U,\Theta,P}(u, \theta, \rho), \tag{4.17}$$

where $f_{U,\Theta,P}(u, \theta, \rho)$ is the joint pdf of wind speed, air density and wind direction. $e(u, \theta, \rho)$ depicts how wind energy is distributed at different values of air density, wind speed and wind direction per unit time and swept area; see also Morrissey et al. (2010) and Carrillo et al. (2014). Assuming that air density is constant, Eq. (4.17) is simplified to

$$e(u, \theta; \rho) = \frac{1}{2}\rho u^3 f_{U,\Theta}(u, \theta). \tag{4.18}$$

The total wind power density can be obtained from the above relation by

$$E(u, \theta; \rho) = \frac{1}{2}\int_{u_{\min}}^{u_{\max}} \int_0^{2\pi} \rho u^3 f_{U,\Theta}(u, \theta)\, \mathrm{d}u\mathrm{d}\theta, \tag{4.19}$$

where $u_{\min}$, $u_{\max}$ are the values of the minimum and maximum observed wind speed in the area, respectively. Moreover, the output $E_T(u, \theta)$ from a single wind turbine, with a power curve P_T is obtained by

$$E(u, \theta; \rho) = T_r \int_{u_1}^{u_2} \int_{\theta_1}^{\theta_2} P_T(u) f_{U,\Theta}(u, \theta)\, \mathrm{d}u\mathrm{d}\theta, \tag{4.20}$$

where $[u_1, u_2]$ is the effective range of wind speed for the particular turbine, $[\theta_1, \theta_2]$ is the domain of wind directions at the particular site, and T_r is the particular time period considered.

From Eq. (4.18) very useful and detailed characteristics of the available wind resource can be obtained at an area that can be further used as design parameters for the wind turbines and wind farm layout; see also Carta et al. (2008b). In this connection, in Figure 4-23(a), the wind power density per wind speed for five different and characteristic wind directions, i.e. $e(u, \theta = \theta_0; \rho)$, is depicted for Mar de Alboran.
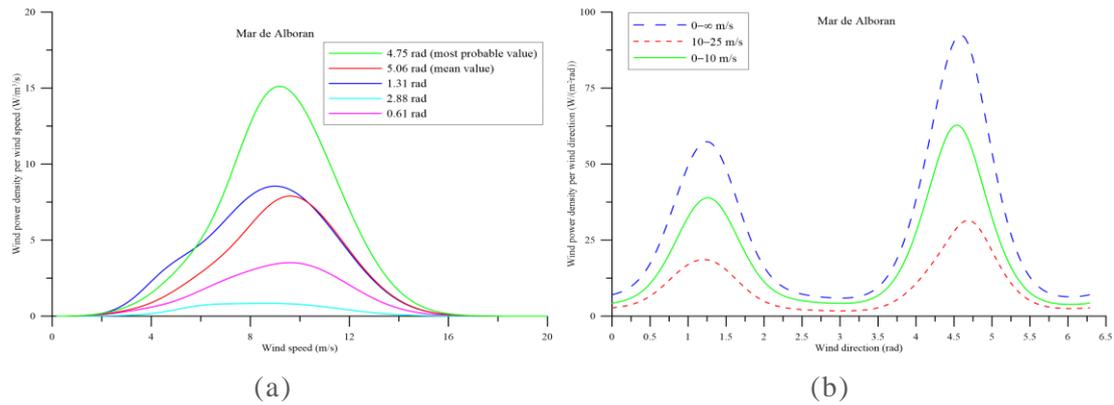
Figure 4-23. Wind power density at various wind directions as a function of (a) wind speed and (b) wind direction.

Note that Mar de Alboran, the wind power density takes its maximum value for the prevailing wind direction, which is the typical parameter for setting the wind farm layout (Ng and Ran, 2016). However, there can be cases when the maximum of the wind power density curve is obtained for the mean wind direction, suggesting that the most probable value of wind direction is not necessarily the one that provides the maximum wind power density. In Figure 4-23(b), the wind power density per wind direction for three wind speed ranges is provided, i.e. $\int_{u_1}^{u_2} e(u, \theta; \rho) \mathrm{d}u$, where $[u_1, u_2]$ is the particular wind speed range considered. Evidently, the range $[0, \infty)$ for wind speed provides the largest values of wind power density. Note that if the horizontal axis in the latter figure is expressed in degrees then the values of the depicted distribution of wind power density will change.

The above presented methodology can be directly and equally well applied for estimating the bivariate pdf of wind speed and direction at the turbine hub height. Combining this information with the turbine characteristics, the annual energy production of a wind farm can be estimated; see also Chowdhury et al. (2013).

4.4.6　Final comments and discussion

The obtained results provided consistently the best values with respect to the above statistical measures for the JW model, suggesting more appropriate fits at the two examined locations compared to the other models. Therefore, it can be concluded that the JW family may be used as a solid base for bivariate modelling of wind speed and direction (compared to the FGM and PLA families). Moreover, in the Greek location, the best bivariate fit for each examined family was obtained with respect to marginal mixture distributions for wind speed. On the other hand, the best-fit wind speed distributions did not provide in general the best-fit models in the bivariate case. In this respect, it seems that the consideration of several different (and not only the best-fit) marginal wind speed distributions is necessary for obtaining the best bivariate fit. FGM and PLA families provided values of the proposed statistical measures very close to each other, with the former one performing better than the latter one in the majority of cases. After obtaining the best bivariate model for the Spanish location, it was applied for estimating wind power density per wind speed and per wind direction as a real-world scenario in wind farm design.

Nevertheless, further research in bivariate modelling of wind speed and direction seems to be essential. A step towards this direction may consist in the implementation of suitable models from the fields of bivariate copulas and kernel density functions. Furthermore, the consideration

of alternative circular models for the modelling of $f_\Psi(\psi)$ and/or $f_\Theta(\theta)$ (e.g. through non-negative trigonometric sums) in the JW model may provide more detail in the dependence structure of wind speed and wind direction and more accuracy in the representation of the bivariate family.

4.4.7 Recent advances

The superiority of JW has also been proved in another analysis, whose results have not been published yet, that deals with the simultaneous study of wave energy flux and mean wave direction by means of parametric and non-parametric bivariate distributions in the Greek Seas. This study stresses the importance of including the directional wave parameter in wave energy resource assessment studies through a straightforward application from the methodology presented. The joint description of these two variables have not been presented yet in relevant studies, although it is of high importance for the emerging wave energy sector. Let us note that it is not among the scopes of this work to examine wave energy potential in Greece rather than propose a methodology that should be employed in wave energy resource assessment studies, especially when the performance of the wave energy converter is dependent on wave direction (e.g. the Pelamis device).

In the context of this work, numerous parametric models have been addressed, either conventional or mixture, and two non-symmetric kdf (Gamma and Lognormal kdfs) for the linear variable in the univariate case. The directional variable is adequately described with one parametric pdf (a finite mixture of the von Mises distribution) and one kdf (the Wrapped Normal kernel model). After a thorough evaluation similar to the one presented in Section 4.4.3, the best univariate (parametric) models for the linear variable are selected to proceed in the bivariate case. The construction of the bivariate distribution functions of the examined parameters (i.e. wave energy flux and mean wave direction) is accomplished through three families of distributions in the parametric case, i.e. FGM and PLA families of distributions and JW model, and the multiplicative kdf in the non-parametric case. A common feature to all these bivariate models (parametric and non-parametric) is that their density functions rely on the corresponding univariate marginal distributions, which are known beforehand (coming from the marginal data) while the parametric bivariate models rely as well on an additional parameter that quantifies the correlation/dependence of the variables. A detailed evaluation of the resulting bivariate distributions is made by applying six bin-defined statistical metrics.

Indicatively, some preliminary results from this analysis are presented for Zakynthos location. In Figure 4-24, the histograms of wave energy flux and wave direction are shown along with the corresponding univariate distributions. The two kdfs and the WW model present a very similar behaviour as regards the linear variable while the two fitted directional models seem to be close to each other with small deviations regarding the peaks at 70° and 285° with the Wrapped Normal kdf underestimating the histogram in both cases.

The values of the evaluation metrics for the bivariate parametric and non-parametric distributions are provided for Zakynthos buoy in Table 4-20. Boldface numbers denote the best value obtained from each metric for the examined bivariate families. For three out of six evaluation metrics Lognormal kdf is the best model compared to both parametric and non-parametric models and for three out of the six evaluation metrics the JW model provides the best fit (but for two different parametric models).

(a)   (b)
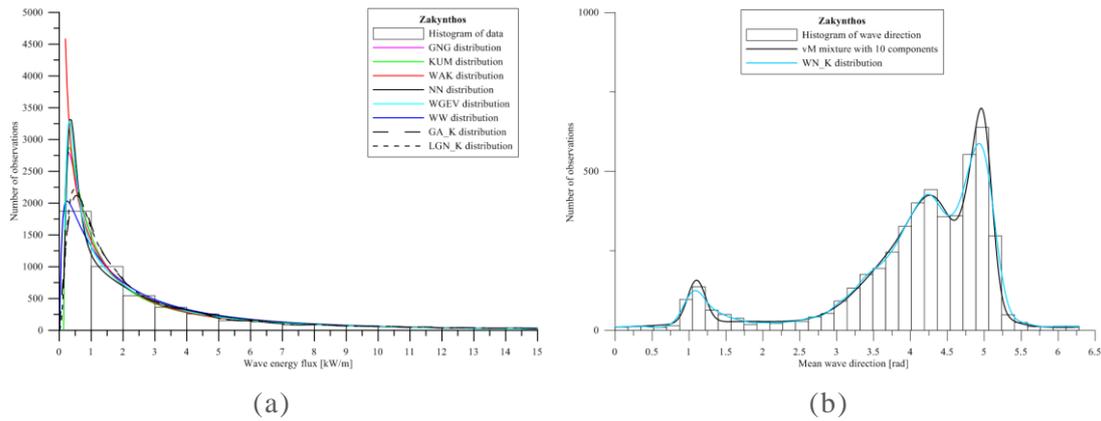
Figure 4-24. Histograms of (a) wave energy flux along with the best fits of the parametric and non-parametric linear distributions and (b) wave direction along with the fitted circular distributions for Zakynthos.

To sum up, the results from both studies demonstrate the significance in the use of parametric univariate or bivariate models that are characterized by simplicity and straightforward implementation and can capture almost all the information of a data set into just a small number of parameters. Furthermore, the inconsistency between univariate and bivariate distributions confirms the value of the proposed methodology in ocean energy assessment studies.

Table 4-20. Values of the goodness-of-fit criteria for three bivariate families of two parametric models and the bivariate kernel models for Zakynthos.

Evaluation measures	NN			WW			GAM kdf	LGN kdf
	JW	FGM	PLA	JW	FGM	PLA		
IA	**0.994**	0.981	0.993	**0.994**	0.982	0.992	0.992	0.993
MAE	0.366	0.487	0.401	0.371	0.489	0.409	0.275	**0.269**
$R_{a,2}^2$	**0.979**	0.930	0.972	0.977	0.931	0.970	0.976	0.976
RMSE	**1.533**	2.785	1.773	1.591	2.757	1.848	1.742	1.647
RRMSE	0.243	0.245	0.245	0.226	0.227	0.222	0.177	**0.176**
χ^2	273	492	372	251	466	335	169	**162**

4.5 Extreme modelling of metocean parameters including directionality

4.5.1 Synopsis

A wide range of wave energy applications rely on the accurate estimation of extreme wave conditions while some of them are frequently affected by directionality. In this section, four offshore/nearshore locations in the eastern Mediterranean Sea are selected with relatively high wave energy flux values and extreme wave heights are examined with wave direction as a covariate. The GP distribution is used to model the extreme values of wave height over a pre-defined constant threshold, with its parameters being expressed as a function of wave direction through a smooth form of Fourier series. In order to be consistent with the analysis obtained from the independent fits for the eight directional sectors of 45-degree width, the estimation of parameters is based on a penalized maximum likelihood criterion that ensures a good agreement between the two approaches. The obtained results validate the integration of directionality in extreme value models for the examined locations, and design values of significant wave height are provided with respect to direction for the 50- and 100-year return period with bootstrap confidence intervals.

This analysis has been accepted for publication in:

Karathanasi, F., Belibassakis, K., Soukissian, T. **Directional extreme value models in wave energy applications**. *Atmosphere*, in press.

4.5.2 Data and numerical results

Reanalysis wave data from the ERA-Interim database for four grid points located in the Eastern Mediterranean Sea were used. The locations were selected according to their potential for development of wave energy projects due to their high estimates of wave energy flux; see for example, Ayat (2013); Karathanasi et al. (2015); Besio et al. (2016); Emmanouil et al. (2016). The wave parameters that were obtained for the purposes of this work were the significant wave height H_S and the mean wave direction θ_W. The geographical coordinates, the measurement period and the sample size of each grid point are listed in Table 4-21.

Table 4-21. Station names, geographical locations, recording periods and sample sizes of wave time series.

Buoy name	Latitude, Longitude (°)	Period	Sample size
Aegean Sea	37.75°N, 25.25°E		
Ligurian Sea	43.25°N, 9.75°E	1979–2014	52596
Otranto Str.	40.25°N, 19.00°E		
Sicily Str.	37.75°N, 12.25°E		

The results for the basic statistical parameters for H_S are presented in Table 4-22. The grid point at Sicily Strait (hereafter, called 'Str.') seems to be a location with intense waves and moderate variability followed by that in the Aegean Sea, which has values of skewness Sk_{H_S} and kurtosis Ku_{H_S} closer to zero denoting a less asymmetric dataset. On the other hand, the grid points at the Ligurian Sea and Otranto Str. present higher variability and lower mean and median values.

Table 4-22. Basic (linear) statistics of significant wave height H_S at the examined locations.

Location	m_{H_S} (m)	med_{H_S} (m)	min_{H_S} (m)	max_{H_S} (m)	s_{H_S} (m)	CV_{H_S} (%)	Sk_{H_S} (–)	Ku_{H_S} (–)
Aegean Sea	1.0	0.8	0.1	5.4	0.7	69.5	1.3	5.3
Ligurian Sea	0.6	0.5	0.1	5.4	0.5	80	1.8	7.6
Otranto Str.	0.5	0.3	0.0	3.8	0.4	85.5	1.9	7.7
Sicily Str.	1.0	0.8	0.1	6.4	0.7	74.4	1.7	7.1

In Table 4-23, the values of the basic circular parameters for θ_W are provided. The rather high value of Ku_{θ_W} for the Aegean Sea (0.5) denotes a rather peaked distribution of wave directions, which is also confirmed by the low value of s_{θ_W} (1.1). The low values of $\overline{R}_{\theta_W}$ denote a weak concentration of data about the mean direction.

Table 4-23. Basic (circular) statistics of wave direction at the examined locations.

Location	m_{θ_W} (deg)	$\overline{R}_{\theta_W}$ (–)	V_{θ_W} (–)	s_{θ_W} (–)	Sk_{θ_W} (–)	Ku_{θ_W} (–)
Aegean Sea	353.5	0.4	0.6	1.1	0.4	0.5
Ligurian Sea	272.3	0.3	0.7	1.2	−0.2	0.2
Otranto Str.	240.1	0.2	0.8	1.3	−0.3	−0.3
Sicily Str.	287.5	0.4	0.6	1.1	0.3	0.2

For each examined location, seven different combinations of the methods for threshold selection and declustering are performed, i.e., each of the threshold selection methods (mean

excess function, threshold stability plot, percentile) is combined with runs and intervals declustering methods along with DeCA declustering method, from which the threshold is obtained as the median of the declustered values. Firstly, the threshold values of H_S are estimated irrespective of Θ_W. After a sensitivity analysis, the 95th percentile was used to derive threshold values, since higher percentiles provided a smaller sample of extreme data that result in large variance. As regards threshold values from mean excess and threshold stability plots, the packages 'evmix v2.11' and 'extRemes v2.0.10' in R were used, respectively; the corresponding graphs are provided in Figure 4-25. In Table 4-24, the threshold values of H_S for each location and method are summarized. The maximum threshold values are systematically provided by the DeCA method while the minimum ones from the mean excess. It is obvious from the mean excess plots of all locations that the decreasing behaviour of the mean excess function shows that the higher we go in the sample data, the lower the excess values are, indicating a thin-tailed behaviour of the distribution.

Table 4-24. Threshold values of significant wave height by threshold selection method for the examined locations.

Threshold selection method	Aegean Sea	Ligurian Sea	Otranto Str.	Sicily Str.
95th percentile	2.32	1.62	1.24	2.47
Mean excess function	1.90	1.30	0.96	2.00
Threshold stability	2.10	1.50	1.00	2.10
DeCA	2.61	1.89	1.25	2.66

(a)

(b)

(c)

(d)
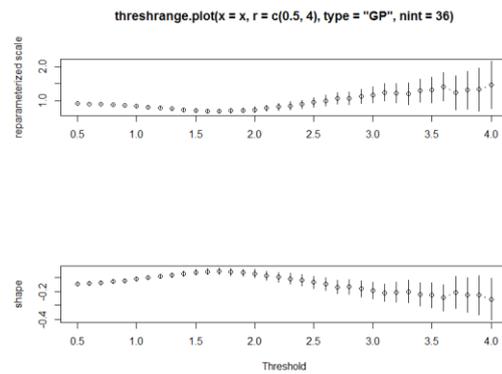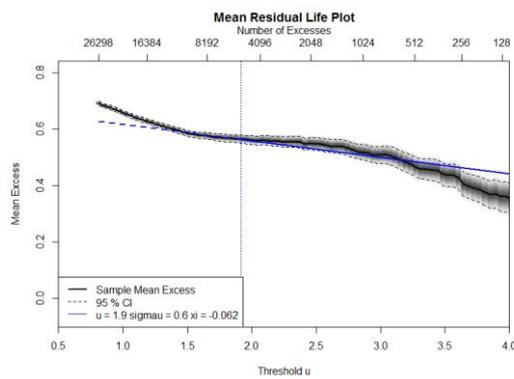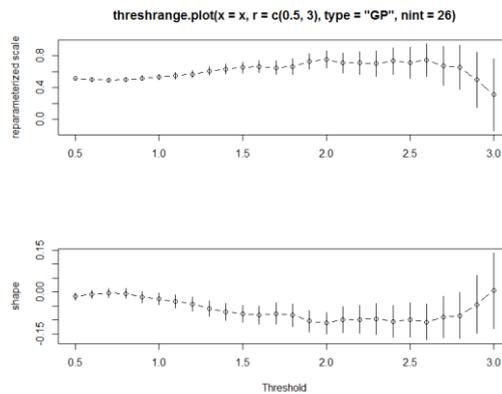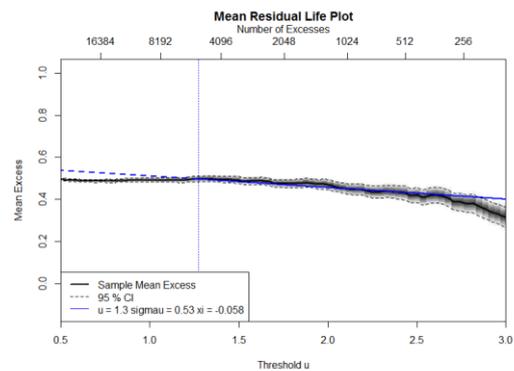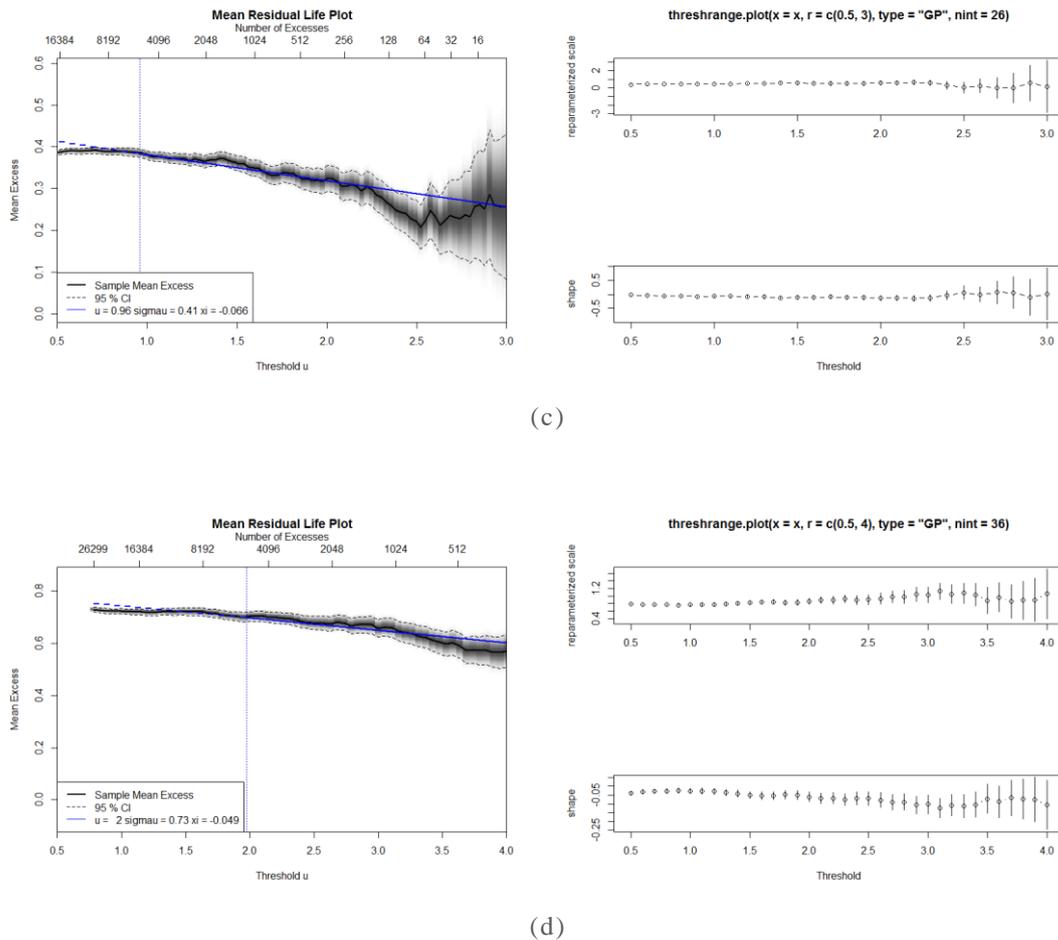
Figure 4-25. Plots of mean excess function (left panels) and threshold stability (right panels) for (a) Aegean Sea, (b) Ligurian Sea, (c) Otranto Str., and (d) Sicily Str.

In Table 4-25, the number of exceedances for H_S after implementing the declustering methods for each threshold selection method is provided for the examined locations. These H_S exceedances along with the corresponding values of Θ_W are used for fitting the directional extreme value model described in Section 3.7. Let it be noted that for runs declustering, a run length of 36h was chosen as the most representative for the examined locations, providing sufficient data for the subsequent analysis. Mean excess function and intervals declustering method provide systematically the largest number of exceedances for all locations. On the other hand, DeCA provides the smallest one, rendering its position disadvantageous in the directional extreme value analysis, since a sufficient number of exceedances (>20 (H_S, Θ_W) pairs of extreme values) is preferred for each 45-sector in order to obtain reliable results from the GP distribution fit.

With the final exceedances in hand, the LR test was performed to determine the order of the Fourier model that sufficiently describes the variability of the extreme value parameters for each location. As shown in Table 4-26, the majority of the considered combinations of methods for threshold selection and declustering for the examined locations concerns the first order Fourier model apart from Ligurian Sea, where the higher order model indicates its directional complexity. Let it be noted that the initial values for the ML approach are obtained by estimating the parameters of the Fourier model from the independent fits by least squares method, which implies a sufficient number of equations according to the number of the unknown parameters (i.e. the order of the Fourier model). Thus, in order to ensure a fair

comparison between the combinations of the abovementioned methods, when the number of the 45-width sectors with sufficient number of exceedances (>20) was less than three (out of eight) for the first order Fourier model, the corresponding combination of methods was omitted from the analysis. The restriction for the second and third order models is five and seven sectors, respectively. The results of Table 4-26 in italics denote the combinations of methods that satisfy these two restrictions. DeCA method is not included henceforth because even for the first order model, the sectors satisfying the above conditions was less than three.

In the estimation of parameters with the penalized ML, an additional constant w needs to be determined. This constant is estimated based on the minimum value of mean absolute error between the estimated parameters from the directional extreme model and the ones obtained by the independent fits from the successive directional sectors of 45-degree width, provided simultaneously for both extreme parameters ξ and σ_u. The obtained results are shown within the parenthesis in Table 4-26.

In this part of the analysis, the standard directional extreme value model was also considered (i.e. when the parameters are estimated without the consideration of the penalty term) in order to verify the use of the penalized factor in the directional model for the estimation of the parameters. Two examples are provided in Figure 4-26 for Ligurian Sea and Otranto Str. locations considering different combinations of methods, a different order for the Fourier model and different weighting constants w. The solid line corresponds to the standard directional model (i.e. $w = 0$), the dashed line corresponds to the directional model, the parameters of which were estimated using the penalized ML with the corresponding weighting constant w (see also Table 4-26) and the circles correspond to the estimates of the parameters obtained from independent fits with data from successive sectors (with width 45°) that are assumed to follow a GP distribution. From this figure it is shown that the estimates obtained from the penalized ML provide systematically better results than the standard method when compared with the estimates derived from the fits of successive sectors, even for a small weighting constant.

Table 4-25. Number of exceedances of significant wave height for each combination of methods and for all locations.

Threshold selection method	Declustering method	Aegean Sea	Ligurian Sea	Otranto Str.	Sicily Str.
95th percentile	Runs	323	340	297	288
	Intervals	671	830	782	669
Mean excess function	Runs	383	374	326	328
	Intervals	1234	1303	1229	1064
Threshold stability	Runs	365	359	325	322
	Intervals	939	991	1165	963
DeCA	DeCA	197	285	308	233

Table 4-26. Order of the Fourier model and value of the weighting constant w (within parenthesis) for each combination of methods and for all locations.

Threshold selection method	Declustering method	Aegean Sea	Ligurian Sea	Otranto Str.	Sicily Str.
95th percentile	Runs	*1 (0.20)*	*3 (0.24)*	*1 (0.13)*	*1 (0.06)*
	Intervals	*1 (0.03)*	*3 (0.18)*	*1 (0.12)*	*1 (0.01)*
Mean excess function	Runs	*1 (0.31)*	*2 (0.42)*	*1 (0.22)*	*1 (0.10)*
	Intervals	*2 (0.09)*	*1 (0.17)*	*3 (0.03)*	*1 (0.02)*
Threshold stability	Runs	*1 (0.17)*	3 (0.42)	*1 (0.17)*	*1 (0.29)*
	Intervals	*1 (0.02)*	*1 (0.30)*	*1 (0.03)*	3 (0.03)
DeCA	DeCA	*1 (0.20)*	*3 (0.24)*	*1 (0.13)*	*1 (0.20)*
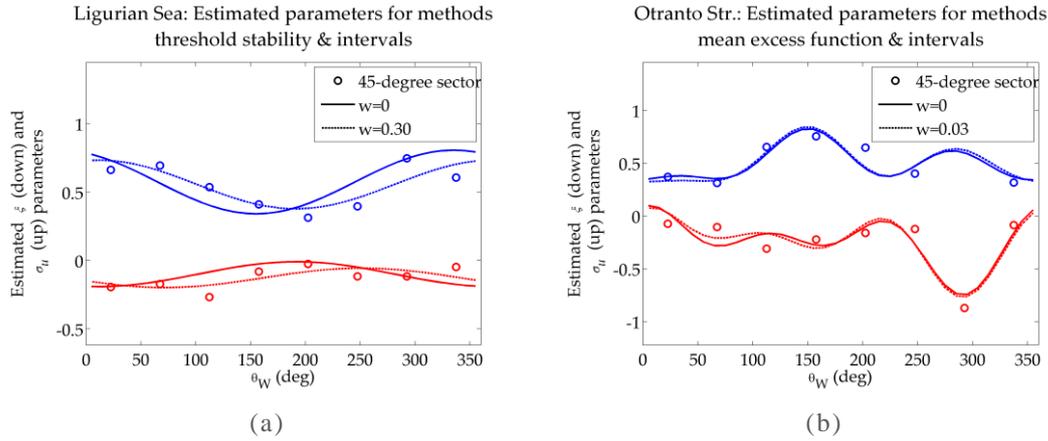
Figure 4-26. Estimated parameters ξ and σ_u of the directional extreme value model obtained with the consideration of the penalty term (dashed line) and without (solid line) for (a) Ligurian Sea, and (b) Otranto Str. Circles represent the estimates from the independent fits of the 45-degree sectors.

From these preliminary results in the selected locations, it is evident that both the use of the directional extreme value model and the inclusion of the penalty term in ML method are essential for the reliable estimation of the design values of H_S and the confidence intervals.

We proceed with the estimation of the design values of H_S for 50- and 100-year return period for the combination of methods that provide the largest sample size of exceedances, i.e. the mean excess function for threshold selection and the intervals declustering method. In Table 4-27, the values of the estimates and the corresponding 95% confidence intervals using the BCA bootstrap method, with number of bootstrap samples $R = 2000$, are given for all locations. As was concluded Coles and Simiu (2003), bootstrapping can provide reliable and realistic estimates for uncertainties in extreme value analysis if carefully implemented.

Figure 4-27 shows H_S design values with direction for the 50- and 100-year return period by considering three different approaches; the blue solid line represents the estimates from the proposed directional model, the green dashed line represents the estimate obtained by the GP distribution without the consideration of the directional complexity of its parameters (omni-directional case) and the red circles represent the estimates from the independent fits of the eight consecutive directional sectors. In order to assure consistency between the results from the omni-directional case and the independent fits from each directional sector, the return period is multiplied by the number of sectors as was suggested by Forristall (2004). In this way, the product of the probabilities obtained from each sector equals the probability of non-exceedance from the omni-directional criterion. For all locations, the design value obtained from the standard GP fit is lower compared to the estimates provided at the peaks of the directional model. Moreover, the design values estimated by the sectors with the largest number of observations are always higher than the corresponding design value obtained from the standard GP fit. The performance of the proposed directional model is apparently satisfactory for Aegean Sea and Otrantro Str. (Figures 4–27(a) and (b), respectively) while for the rest locations the model has relatively large deviations from the independent fits at particular directional sectors; see, e.g. the south-western directional sector of Figure 4-27(d). A possible explanation could be the low order of the Fourier model; see also lower left panels of Figure 4-28(b) and (d), where the range of the lower bounds of confidence intervals is relatively high.

Table 4-27. Point and interval estimates of the directional model for all locations.

Parameter	Aegean Sea	Ligurian Sea	Otranto Str.	Sicily Str.
A_{10}	−0.17	−0.07	−0.24	0.00
	(−0.55, −0.06)	(−0.25, −0.03)	(−0.58, −0.15)	(−0.40, 0.06)
A_{11}	0.10	−0.02	0.00	−0.16
	(−0.28, 0.18)	(−0.20, 0.03)	(−0.25, 0.06)	(−0.79, −0.07)
A_{21}	−0.20	0.07	0.16	−0.02
	(−0.54, −0.11)	(−0.13, 0.14)	(−0.35, 0.28)	(−0.57, 0.08)
A_{12}	0.14		0.16	
	(−0.05, 0.21)		(−0.30, 0.24)	
A_{22}	0.06		0.15	
	(−0.42, 0.23)		(−0.25, 0.25)	
A_{13}			0.15	
			(−0.38, 0.28)	
A_{23}			−0.07	
			(−0.28, −0.02)	
B_{10}	0.66	0.54	0.50	0.71
	(0.36, 0.74)	(0.46, 0.56)	(0.26, 0.54)	(0.27, 0.74)
B_{11}	−0.02	0.17	−0.14	0.37
	(−0.27, 0.05)	(0.03, 0.21)	(−0.31, −0.10)	(−0. 45, 0.44)
B_{21}	0.35	−0.06	−0.01	−0.09
	(−0.13, 0.48)	(−0.18, 0.00)	(−0.21, 0.06)	(−0.96, −0.01)
B_{12}	−0.13		−0.01	
	(−0.39, −0.07)		(−0.21, 0.05)	
B_{22}	0.00		−0.14	
	(−0.57, 0.15)		(−0.27, −0.06)	
B_{13}			−0.03	
			(−0.23, 0.03)	
B_{23}			0.09	
			(−0.07, 0.13)	

In the upper panels of Figure 4-28, the wave rose diagrams of H_S and Θ_W representing the exceedances obtained from the implementation of mean excess function for threshold selection and the intervals declustering method are presented for all locations. In the lower panels of Figure 4-28, the 50- and 100-year H_S design values are shown along the 95% confidence intervals estimated by the BCA method. These levels seem reasonable when considering that the expected lifetime of wave energy converters is 20–25 years on average. A general remark concerning all locations is that the range between the H_S design value and the upper bounds is much smaller than the corresponding range with the lower bounds. Another remarkable result is that in two locations it is not expected to encounter extreme H_S values from the dominant directional sector but from the next one, which may have a more limited amount of data. Since the results of the 50- and 100-year return period are similar, the following comments can be summarized for both return periods per location:

- For Aegean Sea, the dominant sector for extreme wave heights is the northern one, probably attributed to the Etesian winds, which gives extreme values up to 7m at this sector, and lower values characterizes the rest directional sectors (e.g., for the sector $[50°, 310°]$ the H_S value is 4.3m in the mean) as regards the 50-year return period. Furthermore, the low values of the lower bound of the 95% confidence intervals in the north-western sector can be justified by the lack of data obtained from the implementation of the specific combination of methods.
- For Ligurian Sea, the north-eastern sector is characterized by high values of H_S (5.4m for the 50-year return period), even though it is the second dominant directional sector for H_S, while the southern sector, with the least amount of extreme data, provides the lowest values (3.6m).
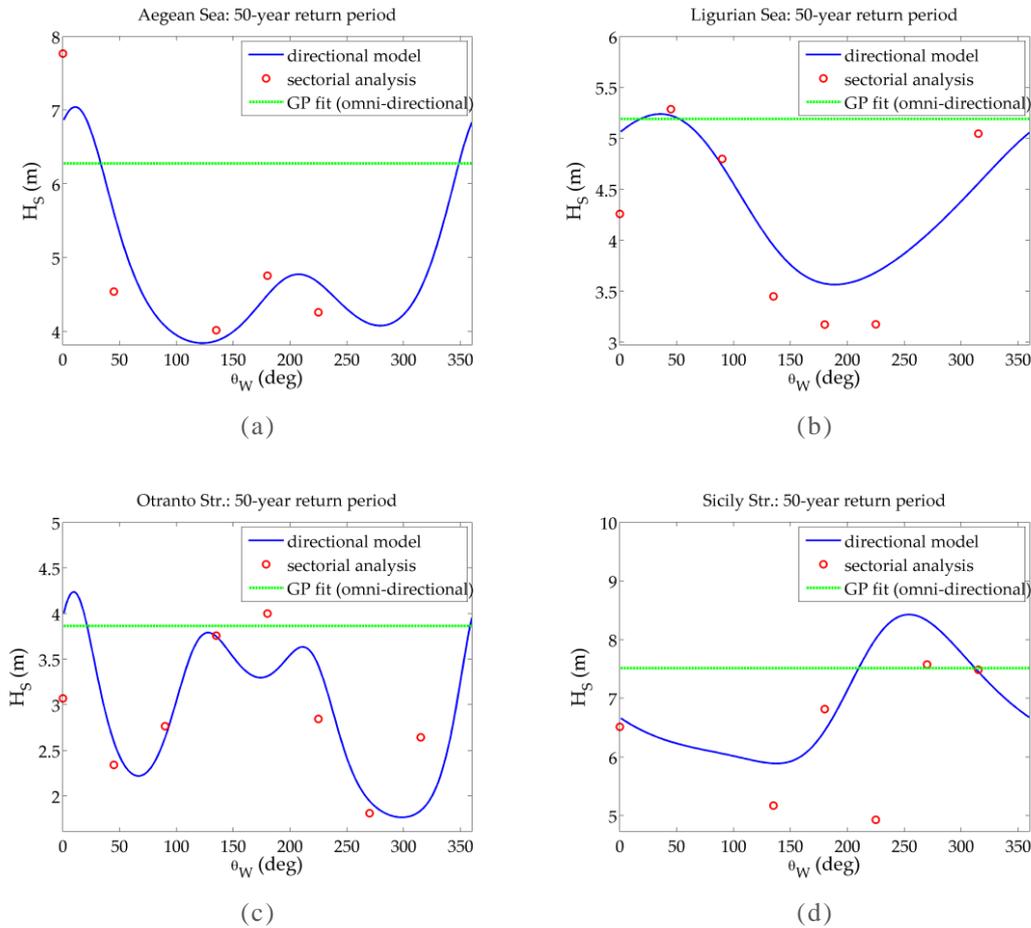
Figure 4-27. H_S design values for the 50-year return period obtained by the proposed directional model (blue solid line), the GP distribution without the consideration of directionality (green dashed line) and the independent fits for (a) Aegean Sea, (b) Ligurian Sea, (c) Otranto Str., and (d) Sicily Str.

- For Otranto Str., the two dominant wave directions (in the south and south-eastern sectors) are translated in two consecutive peaks in the H_S design value graphs while the two concave forms (in the north-eastern and western sectors) correspond to the sectors with the minimum amount of extreme data. Note that the form of the lower bounds differs from the one of the H_S design value.
- For Sicily Str., the location with the most intense sea states according to the analysed hindcast wave data, the second dominant directional sector for H_S (i.e., the western) is characterised by the highest H_S design values (8.4m for the 50-year return period) and the lowest values are observed for the south-eastern sector (5.9m for the 50-year return period). This location presents the highest uncertainty in the estimation of the design values; the largest difference between the lower bounds and the H_S design value is close to 6.3m for the 50-year return period encountered in the south-western sector.
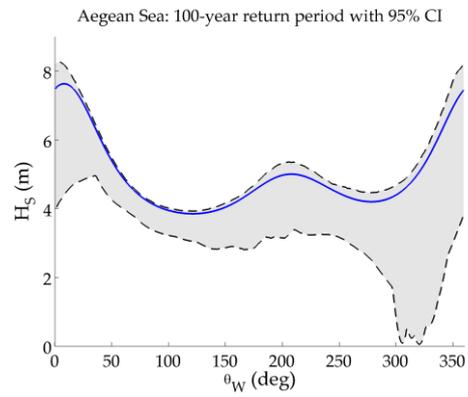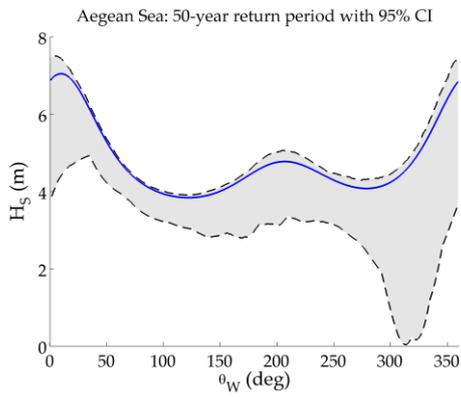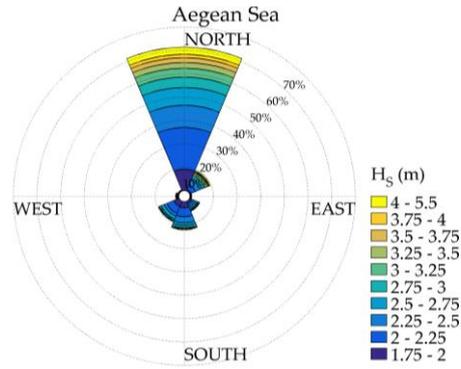
(a)
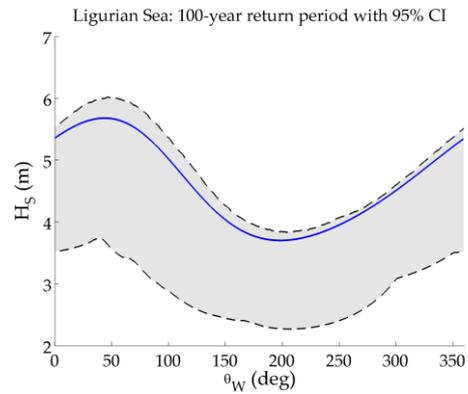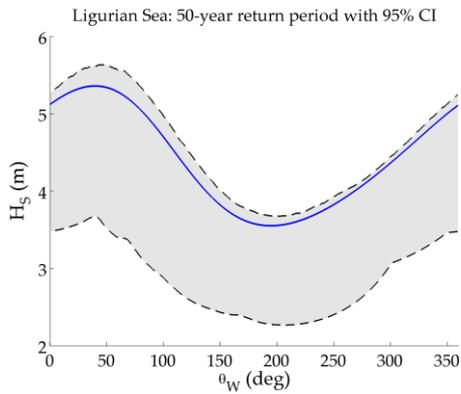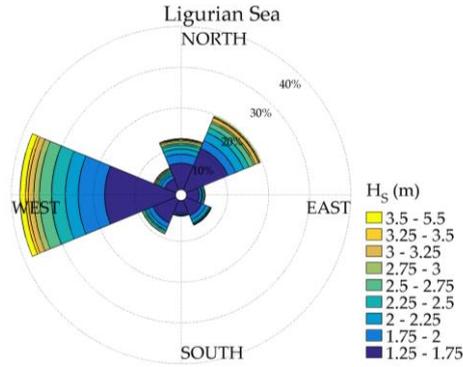
(b)
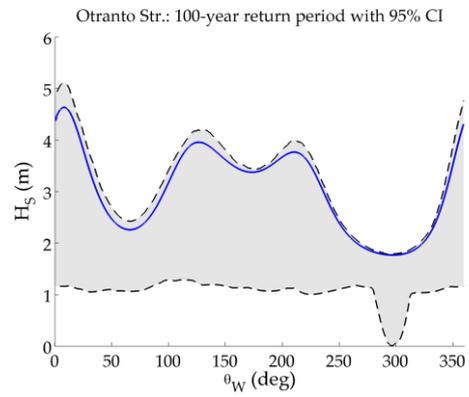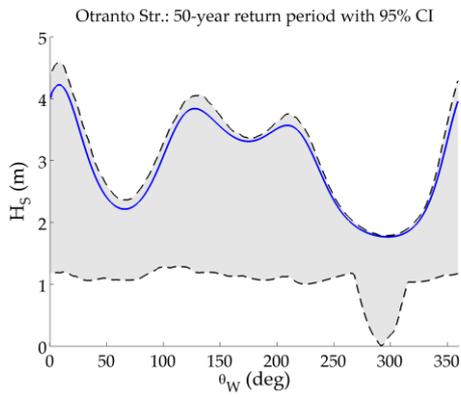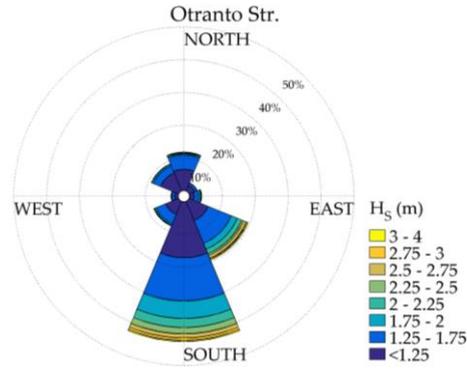
(c)
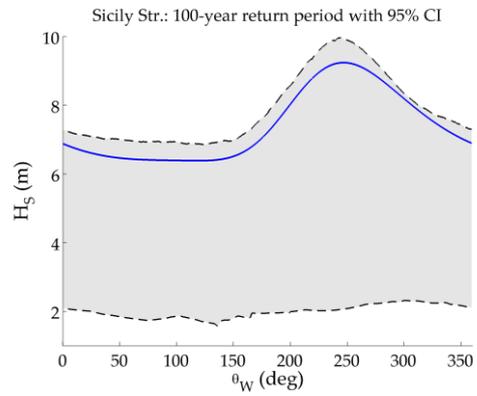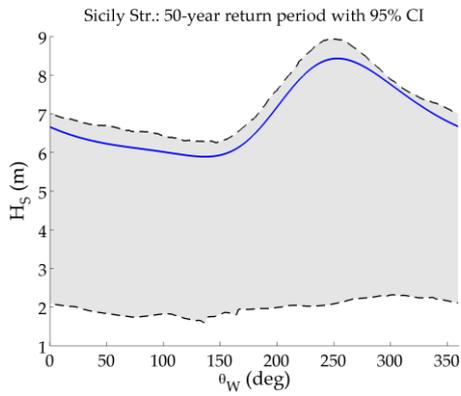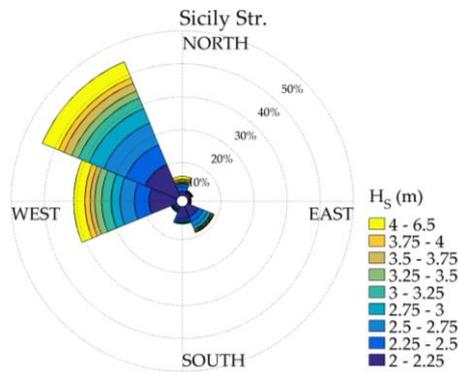
(d)

Figure 4-28. Wave rose of H_S exceedances (upper panel) and H_S design values for 50-year (left column) and 100-year return period (right column) with bootstrap 95% confidence intervals for (a) Aegean Sea, (b) Ligurian Sea, (c) Otranto Str., and (d) Sicily Str.

4.5.3 Final comments

Estimation of design values of wave parameters by means of directional extreme value models can be in favour of extreme value models that ignore direction in wave energy applications, where the consideration of directionality is crucial in the design of marine structures. With the increasing availability of long-term directional metocean data mainly from numerical models, it is strongly advised to take advantage of directional extreme value models in optimizing the performance and costs of marine facilities.

In this analysis, long-term wave data from four locations in the eastern Mediterranean Sea were analysed. Three threshold selection and two declustering methods were combined to examine the corresponding effect in the determination of the order of the Fourier model and in turn, in the parameter estimates and design values and their uncertainties. After selecting the appropriate threshold for each method for the identification of extreme wave heights and applying the proposed declustering techniques due to the prerequisite of independence, a Fourier form was used to model the parameters of the Generalized Pareto distribution as a smooth function of wave direction. A penalised maximum likelihood was implemented to estimate extreme parameters and ensure consistency with the directionally independent fits. In the majority of the combinations of methods, the first order Fourier series model was found to be adequate for the description of extreme wave heights with direction while higher order models were necessary particularly for locations with more complex directional features, like the location in the Ligurian Sea. Directional design values of significant wave height were provided for the 50- and 100-year period as an objective criterion for design specification purposes and predict reliable extreme wave conditions during the lifetime of a wave energy facility. Confidence intervals of 95% were also provided by the bias-corrected and accelerated bootstrap method. Finally, the present analysis may be useful in other applications related to marine renewable energy sectors, such as the offshore wind sector, and coastal engineering studies (e.g., coastal erosion/accretion studies due to wave action coming from multiple directions).

PART III

APPLICATIONS IN THE COASTAL ENVIRONMENT

Chapter 5 Studying the coastal environment under different time scales

5.1 General

Coastal zones receive a wide range of environmental pressures coming either from natural processes (e.g. sea-level rise, storm surges, hurricanes, etc.) or anthropogenic activities (e.g. fisheries, oil and gas extraction, harbour facilities, tourism, etc.). Adding the increasingly disproportional rates of coastal population density compared to the inland (Neumann et al., 2015), there is an imperative need to manage and protect such areas, as well as human life, effectively. Among the measures that should be taken into account is the forecast of coastal morphological changes that are mainly driven by sediment transport gradients.

The essential properties of coastal morphodynamic processes are the interaction between bathymetry/topography and fluid dynamics (Cowell and Thom, 1994; Dodd et al., 2003) that, on the other hand, are responsible to a great extent for the volume displacement during sediment transport. However, morphological changes depend on the evolutionary nature of all the involved complex processes. Sufficient knowledge of coastal geomorphology, wind and wave climate, and the corresponding complex interaction with sediment particles, and better understanding of all the underlying coastal dynamics in various spatio-temporal scales render coastal evolution more predictable.

A wide research field for the representation of coastal dynamics and morphological evolution is based on deterministic (i.e. process-based) coastal area models, which may include both dimensions in the horizontal plane (2DH model) based on depth-integrated quantities, a vertical profile description added to the 2DH model (quasi-3D model) or the fully three-dimensional equations (3D model), and the representation of the examined processes is computer-based; see, for example, the review of Amoudry and Souza (2011). In a relatively short time and at a low cost, different parameters and scenarios can be applied and tested in the context of an engineering problem but the inherent complexity of the abovementioned dynamic processes renders the development of reliable models a rather demanding task. A lot of research has been carried out in order to develop reliable coupling models, nesting techniques and modelling systems that can scale down the forcing from a large scale (e.g. oceanic waters) to a local one (e.g. coastal waters) so as to predict as accurate as possible sediment transport rates and morphological evolution in coastal areas. Moreover, the understanding of such dynamic mechanisms is crucial not only for the human-oriented activities in the coastal environment and the design and stability of coastal structures but also for the quality of the water by transferring pollutants (Gong et al., 2011) and the ecosystem sustainability of nearshore areas, since sediment contributes to the supply and distribution of nutrients and organic materials (Ikeda et al., 2009).

Among the marine dynamic processes, the significance of winds, and hence waves, stands out mainly due to their structuring nature on the coastal environment in terms of morphological formation and composition. For instance, high-energy ocean events influence, among others, erosion-accretion dynamics by affecting the sediment transport rates of a beach while changes in wave climate (including wave direction) may also affect the sedimentary system (Adams et al., 2011). The degree of severity from the impacts of such an event at a beach depends not only on the characteristics of the event *per se* but also on the characteristics of the beach and the sensitivity of the surrounding ecosystems. Based on the perspective of the frequency and amplitude of waves, two common modelling approaches for the consideration of wave action in sediment transport modelling that can be implemented are the following: the first one deals with the action of individual high waves that collide with the shore for a short time window

(e.g. several hours), and the second one takes into consideration the accumulative action of waves throughout a typical year, with high-energy waves during winter and low-energy waves during summer; see, e.g. Ferreira (2005); Callaghan et al. (2009); Karunarathna et al. (2012); Coco et al. (2014).

The purpose of this chapter is to address the fundamental concepts of wave modelling and wave propagation from the offshore to the coastal areas, define the main features that concern sediment transport and discuss how the two different considerations of time affect the equilibrium of coastal systems. Specifically, in Section 5.2 phase-averaged wave propagation models are briefly discussed and Section 5.3 addresses some general concepts for sediment characteristics which depend on the combined action of waves and currents. In Section 5.4, the first case study is presented with the Varkiza bay being studied under the perspective of episodic events that act for a short time window (hours to some days) and had direct impacts on the coastal topography of the beach. The second case study, discussed in Section 5.5, refers to Sitia bay, and deals with the accumulative wave action, where erosion behaviour is governed by the interaction of storm events and calm periods.

5.2   Modelling wave propagation

Wave action is a dominant factor in the coastal zone by influencing geometry and forming the composition of beaches through currents and sediment transport. When studying coastal morphology, wave transformation, i.e. changes in wave characteristics during wave propagation from the offshore to the nearshore waters, is an essential information since wave data from *in situ* measurements (e.g. from moored buoys) and gridded data sets (global wave models and satellite measurements) are available far from shore. Depending on the spatial and temporal scales, wave evolution can be described by two basic categories of mathematical models. The first one is the phase-resolving models that are based on mass and momentum conservation equations for calculating detailed wave characteristics; mild-slope (Kirby and Dalrymple, 1986) and Boussinesq models (Madsen et al., 1991) belong to this class. These models are computationally expensive thus they are suitable for the wave propagation in shallow waters, where the wave properties vary rapidly, and of limited spatial extent. The second one is the phase averaged (or spectral wave) models that are based on the conservation of the wave action density (Bretherton and Garrett, 1968; Andrews and Mcintyre, 1978) in the presence of currents varying in space, mainly applied for areas (from global to regional spatial scale) in deep water where wave properties vary slowly; see, e.g. Komen et al. (1994); Booij et al. (1999).

The wave action balance equation, which is the governing equation for the latter wave models, can be written in Cartesian co-ordinates as follows:

$$\frac{\partial N}{\partial t} + \frac{\partial}{\partial x}(c_x N) + \frac{\partial}{\partial y}(c_y N) + \frac{\partial}{\partial \sigma}(c_\sigma N) + \frac{\partial}{\partial \theta}(c_\theta N) = \frac{S}{\sigma}, \qquad (5.1)$$

where N is the wave action density, defined as the ratio of the energy density spectrum and relative angular frequency ($N = E/\sigma$), t is the time, c is the propagation velocity in the four-dimensional space x, y (Cartesian coordinates in two horizontal directions), σ and θ (direction of wave propagation), and S is the source term of wave energy balance representing the summation of linear and non-linear interactions. These interactions include the following physical processes that generate, redistribute or dissipate wave energy: wave growth by wind action, wave energy transfer due to non-linear wave-wave interaction and dissipation of wave energy due to white-capping, bottom friction and wave breaking. The solution of Eq. (5.1) provides wave predictions across a computational grid based on the evolution of the wave spectrum.

Among the most popular third generation global wave models that are extensively applied for wave forecasting and the assessment of wave resources in a large spatial scale are the Simulating Waves Nearshore (SWAN) model, developed at the Delft University of Technology (Holthuijsen et al., 1993), Wave Action Model (WAM) model, developed by the WAMDI Group (Group, 1988), WAVEWATCH III (Tolman, 1997) based on WAM model, and MIKE 21 SW of the MIKE 21 suite, developed by the Danish Hydraulics Institute (DHI) International (DHI, 2016).

5.3 Sediment transport: concepts and characteristics

As waves approach the shoreline into shallow water depths, their properties, e.g. wavelength, wave height, period and direction of propagation, are modified significantly and are redistributed due to the varying bathymetry. In particular, within the coastal environment, a part of the wave energy is responsible for the agitation and movement of the bed material (e.g. wave breaking) while wave-induced currents are usually the dominant factor for its transport. For example, relevant studies for the estimation of bed load due to the wave action solely or the combination of waves and currents can be found in Hallermeier (1982); Williams and Rose (2001); Nielsen and Callaghan (2003); Soulsby and Damgaard (2005); Jiang et al. (2015). Due to the continuous response of the sediments to the wave action and currents, the shoreline in turn responds to these physical processes influencing the dynamic equilibrium of the beach; depending also on the seabed and beach material, beach morphology, coastal profile and supply of sediment, the shoreline may be eroded, accreted or stay in equilibrium state.

Generally, sediment transport is divided into two classes:
- longshore transport due to oblique breaking waves generated by longshore currents that moves the sediment parallel to the shore, and;
- cross-shore transport that leads to the onshore or offshore net transport of sediment perpendicular to the shore.

The accurate prediction of seabed level change relies on the accurate estimation of coastal sediment transport since gradients in the sediment transport rates lead to seabed topography changes due to erosion/accretion.

5.3.1 Characteristics of sediment transport

Focusing on the non-cohesive granular sediments in this thesis, the size of the grain is the most important factor for its classification, which can be expressed through statistics (e.g. mean value, standard deviation, skewness, kurtosis) derived from the sand size distribution concerning a particular sand sample or from the measured settling velocity, usually extracted directly from laboratory measurements. For instance, the median particle diameter d_{50} is a representative measure for sand samples, for which half of the sample contains finer particles compared to the other half with coarser ones. The Wentworth scale is the most popular classification of sediment grain by size based on powers of two (Wentworth, 1922), where sand ranges from 0.0625 mm to 2 mm. At this scale, there is also finer-grained sediments referred to as silt and clay, and coarser-grained sediments referred to as gravel (e.g. boulder, pebble). Additional properties include sorting (grading) and shape of grain. The factors that determine the size of sediments are: i) wave energy conditions; ii) sediment sources, and; iii) offshore slope.

In both horizontal and vertical directions of a seabed, characteristics of sediment can vary significantly primarily due to the action of waves and currents. For instance, in the horizontal direction and under high wave conditions, coarse sands are encountered in shallow waters as

finer sediments are deposited in areas with less turbulence (e.g. deep waters) while in the vertical direction, coarse sediments may cover finer ones during intense storm events. Apart from the spatial variations, temporal variations (seasonally or over longer timescales) can also occur.

Note that except for the hydrodynamic/wave conditions of an area and the grain diameter, the pattern of sediment transport is also influenced by the characteristics of the transported material, usually defined by grain (relative and bulk) density, porosity, fall velocity, etc.

The sediment load in a coastal area can be transported in various ways depending on the bed shear velocity. The most commonly modelled ones are:

- bed load transport q_B, where the value of bed shear velocity exceeds the critical value for initiation of movement and the sediment particles are almost continuously in contact with the bed during transport, and;
- suspended load transport q_S, where the value of bed shear velocity exceeds the fall velocity of the particles leading to the lift and suspension of the grains outside the close vicinity of the bed due to the upward impulses carried by turbulent eddies.

The accurate estimation of coastal sediment transport is not a trivial task due to the complexity of the environment and the interdependence of numerous coastal processes in sediment dynamics. In the relevant literature, there is a plethora of theories, assumptions and methods proposed for the study of sediment transport from empirical formulas, which are the most commonly implemented in modelling studies, to more sophisticated experimental techniques. Sediment can be transported under the action of currents, waves and their combination and its movement can be investigated by Lagrangian and Eulerian models; the sediment transport numerical model used in this thesis is based on the latter model. The most controlling factor for the erosion/deposition patterns is the bed shear stress; the total sediment transport, in turn, affects the bathymetry evolution, which results in changes in the current and wave fields. Thus, due to the numerous chain-dependent physical processes involved in sediment dynamics, all these models must be coupled.

5.3.2   Threshold of movement

The sediment movement is dependent on the equilibrium of forces, horizontal and vertical, coming from the water motion, neighbouring grains and hydrodynamic sub-pressure that acts on the surface of the sediments. When the instantaneous fluid forces are just greater than the resisting forces on a particle then this phenomenon is called threshold of sediment motion or sediment incipient motion. In order to predict sediment transport rates and morphological changes at a coastal location, the first step is to predict this threshold.

Both theoretical and experimental studies have been performed for the incipient motion of non-cohesive sediments. The Shields parameter θ is a non-dimensional number widely used as a criterion for the initiation of movement of sediment when the critical bed shear stress of the sediment is exceeded by the shear stress induced by the flow (Shields, 1936). The critical Shields parameter is defined by

$$\theta_{cr} = \frac{\tau_{cr}}{g(\rho_s - \rho)d_{50}},$$

(5.2)

where τ_{cr} is the critical value of the bed shear stress, ρ_s is the density of the sediments and ρ is the corresponding one for the fluid, and d_{50} is the median size of the sediment particle. This stress can be caused either by currents, waves or their combined action.

In case of pure current, the bed shear stress is calculated using simple drag coefficient expressions, which rely either on constant drag coefficients or a logarithmic velocity profile. The latter approach has a greater advantage over the former when morphological changes are considered due to the dependence on the vertical distance from the bed. For a given location at height z above the boundary, the velocity is given by

$$u(z) = \frac{u_*}{\kappa} \ln\left(\frac{z}{z_0}\right), \tag{5.3}$$

where $u_* = \sqrt{\tau_{bc}/\rho}$ is the friction velocity (τ_{bc} is the current-produced shear stress and ρ is the fluid density), κ is the von Kármán (typically equals to 0.41) constant and z_0 is the bed roughness length.

Under the action of pure waves, the critical conditions for sediment motion are expressed through the critical bed shear stress by introducing the concept a friction factor as follows

$$\tau_{bw} = 0.5\rho f_w u_{bw}^2, \tag{5.4}$$

where u_{bw} is the bottom (or near-bed) wave orbital velocity. The wave friction factor f_w depends on the flow regime (e.g. smooth, turbulent), which in turn depends on the wave Reynolds number R_w and the relative roughness $r = \alpha/k_n$ (α is the wave orbital amplitude and k_n is the Nikuradse's bed roughness parameter) of the seabed. Among the variety of formulations that exists for approximating f_w, the most commonly used explicit expressions have been proposed by Swart (1974); Kamphuis (1975); Nielsen (1992).

The representation of bed shear stress developed under the combined action of waves and currents is more complex due to their non-linear interactions. In this case, usually both mean and maximum combined bed shear stresses need to be determined. Several approaches for the parameterization of wave-current interactions have been proposed; see the review of the state-of-the-art knowledge of sediment transport caused by waves and currents by Lu et al. (2015).

5.4 Case study 1: Modelling nearshore hydrodynamics and circulation under the impact of high waves at a coastal area

5.4.1 Motivation

The main motivation of this application is to study the effects of high waves on hydrodynamics and circulation on a sandy beach and, in turn, give insight into their impact on sediment transport processes. Because of the abundance of the available *in situ* measurements, Varkiza coast, in the Saronic Gulf (western Aegean Sea), has been selected as a suitable area for modelling the hydrodynamic and meteorological conditions and estimating sediment transport rates during and after intense sea states/storms by using a quasi-3D sediment transport model based on finite volume method. Specifically, Varkiza coast, located in the homonym bay, forms a part of the north-eastern Saronic Gulf, a semi-enclosed embayment in the south-western Aegean Sea; see Figure 5-1. It is limited in width and length (around 900 m), while at the east side of the coast there is a flume mouth that follows dry/wet epochs. Furthermore, the U-shape and south orientation of the examined coast confine wave action, which is the primary factor for the settlement of sediments. Erosion phenomena are evident due to both the intensive onshore development and physical conditions. The main reasons for choosing the particular coast, apart from the recreational and economic activities that it hosts, refer to the availability of the following features:

-
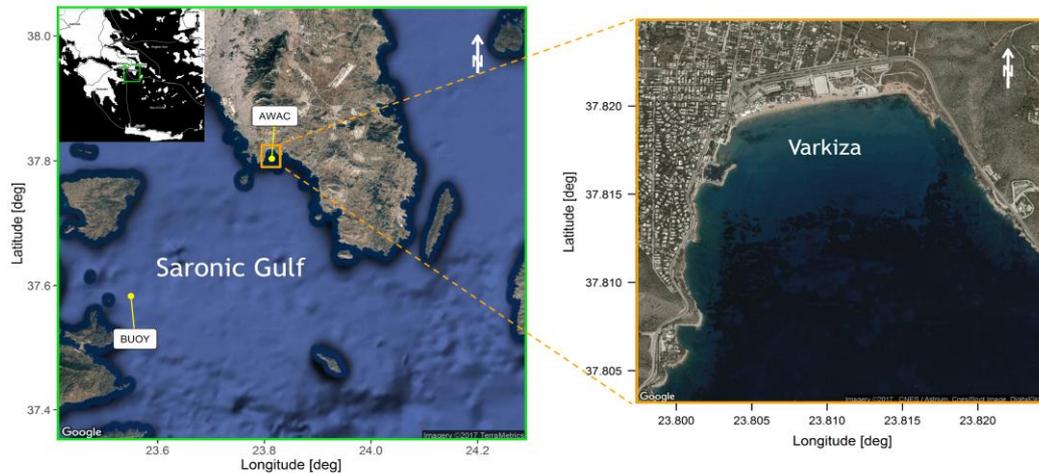
Figure 5-1. Aerial map of the Saronic Gulf along with the locations of the in situ devices (left map), and the study area of the coast of Varkiza (right map) from Google Earth.

- *in situ* measurements as regards the wave parameters from two different data sources; an oceanographic buoy at an offshore location and an acoustic wave and current (AWAC) profiler at the entrance of the bay (see also Figure 5-1);
- a detailed bathymetry up to 25 m water depth inside Varkiza Bay, which was partially based on seabed mapping;
- cross-shore sections along the beach, on which seabed level was measured after intense sea states completing an annual cycle, as well as grain size of sediments (Foteinis, 2014; Skanavis, 2013), and;
- the touristic character of the area, along with the intense socio-economic activities along the beach mainly during summer months, that renders the understanding and prediction of sediment transport phenomenon a critical task.

The results from this analysis have been published in:
Belibassakis, K., Karathanasi, F., 2017. **Modelling nearshore hydrodynamics and circulation under the impact of high waves at the coast of Varkiza in Saronic-Athens Gulf**. *Oceanologia* 59(3): 350–36.

5.4.2   Wind and wave climatology

As regards wind and wave climatology, the analysis was based on a 9-year dataset from an oceanographic buoy that was taken into consideration as a representative location for the examined area. This buoy, deployed at the southern part of the Saronikos Gulf (37.588N—23.558E, water depth ~200 m) belongs to the POSEIDON marine monitoring network that operates under the responsibility of the Hellenic Centre for Marine Research (HCMR) since 2000 (Soukissian et al., 1999). The wind measurements, with reference height 3 m above sea surface, have a 3-h recording interval with 1 Hz sampling frequency (averaged over a 600-second recording period), while the wave measurements have a 3-h recording interval with 1024 s for the sampling period of the free surface. The time series of wind speed and significant wave height is between 08/2007 and 05/2015.

In Figure 5-2(a) and (b), the rose charts of wind speed and significant wave height are presented, respectively, along with the corresponding frequencies of occurrence. From the former figure,

it is illustrated that winds are blowing mainly from the north (sector $[337.5°, 22.5°]$) while, at the same sector, high values of wind speed are also present with the maximum one reaching values up to 17.3 m/s. On the other hand, for the latter figure, the prevailing wave directions (sectors $[0°, 67.5°]$ and $[135°, 157.5°]$) correspond to the less frequent wind directions. Large fetches are evident from the eastern side of the location of the buoy (around 65 km) while in the north and south directions, wave fetch is smaller (15 km on the average). Waves propagating from the west have very low frequency of occurrence, which is reflected also by the very small fetch. Low values of significant wave height (up to 0.5 m) have very high frequency of occurrence (4–5%) coming from the east, while waves characterized with the highest values of the same parameter (up to 3.1 m) propagate from the south-east, attributed to the very large fetch (115 km). Furthermore, note that the scattering of wind directions is broader compared to the wave directions.

As regards water circulation, in Kontoyiannis (2010) direct current observations were analysed at three different time periods and it was concluded that the seasonal flows at the north-eastern part of the Gulf have a northward meandering when north-western, western and southern winds are blowing. Furthermore, the circulation pattern is characterized by a two-layer structure (cyclonic in the upper layer and anticyclonic in the lower layer) from late spring to summer to late fall. In the same work, the time series of current velocity for a 3-month period (11/2003–01/2004) indicated that the currents are in the mean rather weak.

(a)   (b)
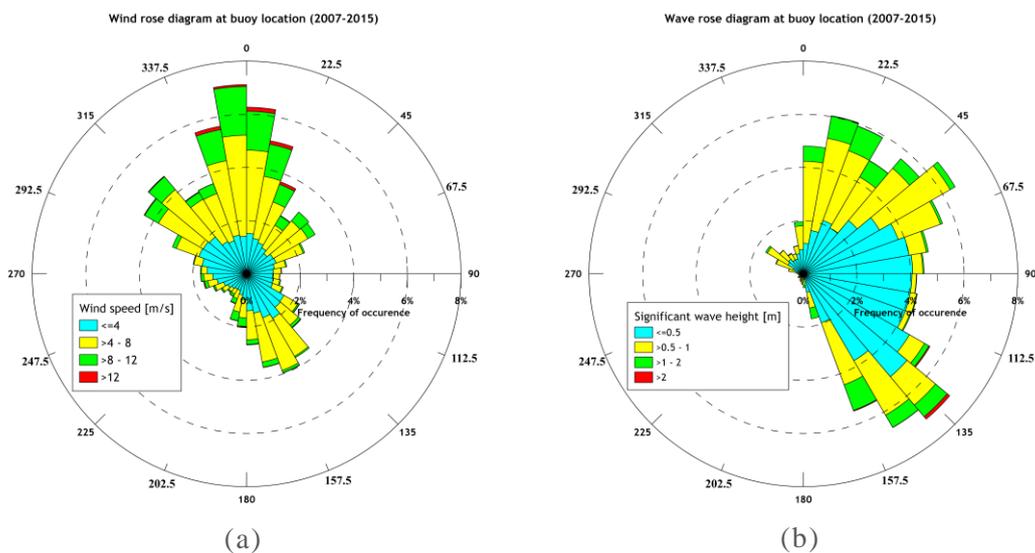
Figure 5-2. Rose diagram of (a) wind speed and wind direction, and (b) significant wave height and wave direction at the buoy location for the period 2007—2015.

5.4.3   Model domain and bathymetric data

In order to manage the computational domain and economize on computation time, the model domain was sectioned into six nested rectangles, going gradually from the outer area (i.e. level 1) up to the entrance of Varkiza Bay (i.e. level 6); see also Figure 5-3(b) for the representation of the different levels. The outer area covers a surface of 45 km x 76 km and the area of Varkiza Bay equals to 2 km x 2.5 km.

In order to determine the variability of flow characteristics in space, model grid resolution is a key factor that affects the quality of the obtained results. The provision of flexible mesh in MIKE21 results to a more accurate representation of the area under study, with the choice of finer mesh elements at local areas of special interest. In this study, various mesh areas were

applied to discretise the water surface, with small triangular elements representing areas where the accuracy in the calculations was important. The final mesh area of the examined area is presented in Figure 5-3(b). The bathymetric information that is necessary for constructing the mesh area for the entire area was obtained by the Hellenic Navy Hydrographic Service (HNHS) from maps of different spatial scales. The bathymetric grid data for the last level (of a 5-m spatial resolution) was obtained by combining a high-resolution map from the HNHS and field measurements provided by the HCMR. In Figure 5-3(a), the 2D bathymetric representation of the examined area is displayed in Google Earth; the deepest water depth is close to 800 m at the south-eastern boundary of the study area.

5.4.4 Input data

The period of the simulation, extending from January 3 to February 19, 2013, was selected so as to include a sequence of extreme events with significant wave heights higher than 2.5 m that were recorded at the entrance of the bay during this period. Furthermore, bathymetry resolution (including flexible mesh) and time step for computations of the HD and SW results are key parameters for the purpose of this study. As concerns the mesh, it becomes progressively finer as we move from level 1 to level 6, which is the local domain at the coastal site of Varkiza. The total number of elements in the whole domain is 12,176, the corresponding number of elements in level 6 is 1600, and the time step is set to $\Delta t = 1800$ s. The latter are found to be enough for numerical convergence of the results concerning the wave quantities that are presented in more detail below. Specifically, numerical investigation shows that the calculated results do not change more than 5% with further enhancement of the mesh at the different subdomains and reduction of the time step.

(a)   (b)
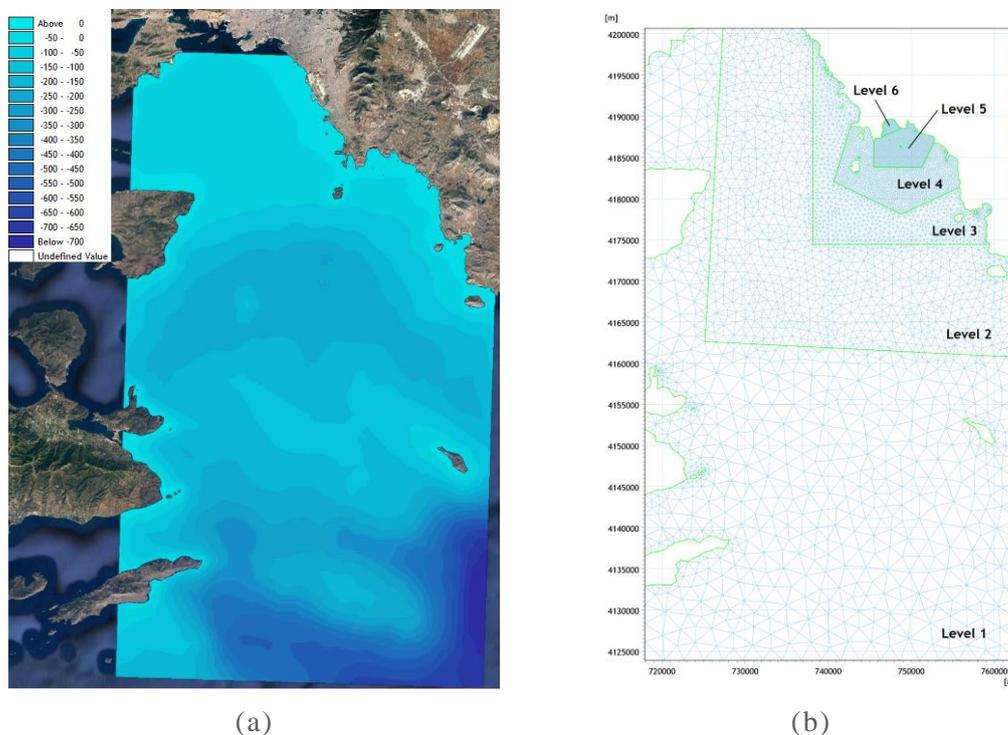
Figure 5-3. (a) The model domain showing the bathymetry of the examined area and (b) the mesh grid for the adopted levels.

The necessary input data for the HD module include the following parameters: wind forcing, radiation stress fields, boundary conditions, atmospheric pressure, bed resistance and eddy viscosity. Eddy viscosity was obtained in the domain from the Smagorinsky formulation with a constant coefficient (equal to 0.28), bed resistance (defined by the Manning number) was only varying at level 6 (with values between 10 and 32 $m^{1/3}$/s), where sediment transport rate is of interest, while salinity and temperature were constant during the simulation (barotropic mode). Regarding boundary conditions, normal fluxes were forced to zero for all variables along both closed and open boundaries, assuming full slip boundary conditions, since all boundaries are far from the area of interest and tidal heights are rather small and do not impact the simulation results. Let us note that tidal heights, predicted from the Global Tide Model Data, were also used as an alternative input for the open boundaries; however, the simulation results were similar to the ones presented in this work.

The effect of the wind forcing on the flow field is included by considering wind speed and wind direction; in this way, wind shear stress is calculated on the water surface. For the numerical simulations, these two variables were considered to be varying in time but constant in domain. Wind speed and direction were obtained by the results of the POSEIDON II weather forecasting system (Korres et al., 2010) that has been developed in the framework of the POSEIDON-II project[8] (2005–2008).

As regards the SW module, the corresponding conditions at the offshore (south) open boundary were varying in space (along the wave generation line) and time. The corresponding input was based on the WAM Cycle-4 code, a third generation wave model, which computes spectra of random short-crested wind-generated waves. The spatial resolution is 1/30° x 1/30° (~3 km) resolving the wave spectrum at each grid point in 24 directional and 30 frequency bins. The wave parameters that were obtained are the significant wave height H_{m_0}, the peak wave period T_P, the mean wave direction θ_{wave} and directional spreading n. The zero upcrossing period T_z, obtained from the WAM model, was converted to the peak wave period T_P by using the following approximate relation (DNV, 2011):

$$\frac{T_z}{T_P} = 0.6673 + 0.05037\gamma - 0.006230\gamma^2 + 0.0003341\gamma^3, \tag{5.5}$$

where γ is the peak enhancement factor of the spectrum. Assuming a JONSWAP spectrum with $\gamma = 3.3$, Eq.(5.5) results in $T_P = 1.2859T_z$.

Other key parameters or coefficients for setting SW module are:
- energy transfer, where quadruplet-wave interaction was considered;
- wave breaking was included by specifying the gamma parameter γ_{wb} (constant in domain, equal to 0.8);
- bottom friction, specified by the Nikurdase roughness k_N (constant in domain, equal to 0.04 m);
- white capping, specified by the two dissipation coefficients (constant in domain) C_{dis}, which controls the overall dissipation rate (set to 4.0), and $DELTA_{dis}$, which controls the weight of the dissipation in the energy spectrum (set to 1.0).

Both wind and wave data were derived from the POSEIDON Live Access Server (LAS, http://poseidon.hcmr.gr/listview.php?id=17), which is a gateway to archived model results, dating from December 1, 2012 to June 30, 2013 with a 6-h time resolution for both datasets. Missing data were filled in by linear interpolation to allow the execution of the simulation; however, interpretation of the simulation results during these time periods should be avoided.

---

[8] The POSEIDON-II weather forecasting system is operational since December 2007 and is applied on a horizontal resolution of 1/20° x 1/20° (~5 km) over the domain covering the whole Mediterranean and Black Sea regions and the surrounding countries.

Finally, regarding the setting up of the ST module, the transport tables have a key role; these tables are built based on all possible combinations according to the specified intervals of the involved parameters (i.e. the root-mean square wave height, peak period, current speed, wave height-to-water depth ratio, angle between current and waves, median grain diameter d_{50} and sediment grading). Additional parameters that are important for this module are forcing from the wave and current action, provided by the HD and SW simulations (see below Figure 5-4 and Figure 5-5, respectively), sediment properties, time step factor (set to 5, i.e. estimation of seabed level and sediment transport every 5th HD time step) and settings for the morphological changes and boundary conditions of the area of interest. In particular, based on measured data, the spatial distribution of the grain diameter of the sediment (d_{50}) over the coastal zone of Varkiza Bay was set to 2 mm for water depths greater than 20 m (where no significant sediment transport is expected). Moreover, in depths less than 3 m there is a gradual increase of d50 from 0.1 mm to 0.4 mm in the S-N direction (moving towards the shoreline), and a variation from 0.35 mm to 0.45 mm in the E-W direction on the shore. Sediment grading was kept constant (equal to 1.45) for the same level.

5.4.5 Model calibration and validation

Model calibration is necessary in order to adjust and improve the agreement between the results of the model simulations and a chosen set of benchmarks (Trucano et al., 2006); in this study, benchmark is a data set obtained by *in situ* measuring devices, which are considered to be the most accurate data sources. On the other hand, validation is the process of verifying that the predictions from the model are consistent with the examined physical events after calibrating the involved parameters or coefficients. Let us note that the data used for the validation should be different from the data used during the calibration phase.

There are numerous parameters and coefficients that should be set so that the model predicts reasonable results; for instance, in the case of the SW model, the parameters that influence the model results regarding the fully spectral formulation and should be adjusted are C_{dis} and $DELTA_{dis}$ dissipation coefficients, gamma and alpha parameters of wave breaking and bottom roughness.

The wave parameters that were used to validate the model were the significant wave height, the zero-crossing wave period and the mean wave direction at the locations where *in situ* measurements were available; two different sources of such measurements were accessible, an AWAC profiler at the entrance of the examined bay and an oceanographic buoy at an offshore location. The latter data source was used to calibrate the wave data input at the offshore boundary of the model domain, where wave data from the WAM model were available, by applying the calibration methods described in Section 2.4.4. Let us remark that in the calibration procedure, more emphasis is given to the significant wave height, since highest waves are expected to have major contribution to the movement of sediments during storm events. For the model assessment, the following statistical measures were applied: RMSE and MAE were used for the linear variables (i.e. significant wave height and wave period) and MCAE and RME for the directional ones (i.e. wave direction); for the corresponding definitions, see Appendix B.4. The validations against both sources of *in situ* measurements showed that there is a good agreement as regards significant wave height and mean wave direction, but wave period exhibits a less accurate performance.

5.4.6 Simulation results

The following results represent the current and wave characteristics and bottom morphology of the examined area for the "extreme" event that occurred on January 18, 2013. The time series of wind speed, wind direction and significant wave height used as input at the offshore boundary are presented. It is evident that southern winds generate the highest values of significant wave height during the simulation period. Moreover, wave height variation is found to be in good agreement with wind speed data, denoting that the waves at this location are mostly wind generated.

HD results

The spatial distribution of current speed and the corresponding direction for the entire area and the coast of interest is depicted in Figure 5-4 during a specific extreme event (on January 18, 2013) that was characterized mainly by south wind and wave directions. The model domain is characterized by low current speeds, of the order of 0.2 m/s. As regards Varkiza Bay, highest values of current speed are observed; locally (at the east side of the bay) current speed reach values up to 0.9 m/s, which is an extreme value encountered very locally during the peak of the storm. The latter high values may be also attributed to the wave direction and the orientation of the coastline. Moreover, in Figure 5-4(b) a counter-clockwise current circulation is evident during this extreme event due to the concave and curvilinear shoreline structure of Varkiza coast and the relatively deep water depths that enhance penetration of waves and currents from easterly sectors. The combination of the above factors produces offshore currents near the western part of the study area. From the analysis, it seems that tidal currents might be of secondary importance in the context of coastal erosion.

SW results

In Figure 5-5, the spatial distribution of the significant wave height and mean wave direction is presented over the model domain. The analysis of the results shows that the significant wave height is reduced as the waves propagate towards the shallower water depths of Varkiza beach; see also Figure 5-5(b). Near the coast the wave height is lower than 2 m with a mean wave period around 7 s (not shown here).

(a)   (b)
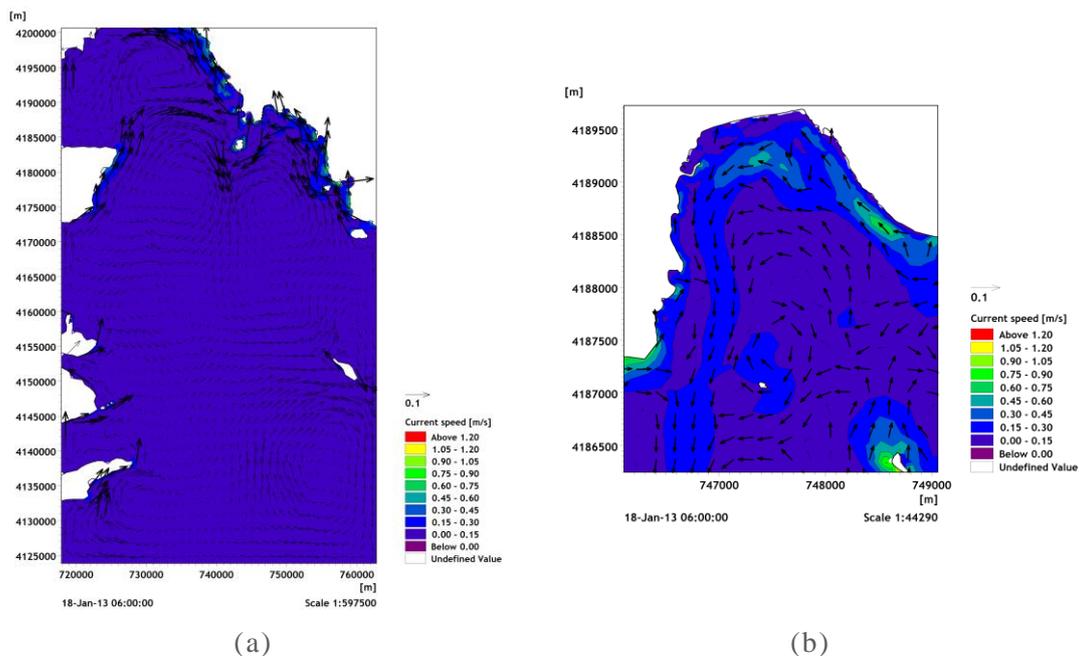
Figure 5-4. Spatial distribution of (a) current speed and current direction for the entire model domain and (b) for Varkiza bay at a specific time step of the simulation.

(a)   (b)
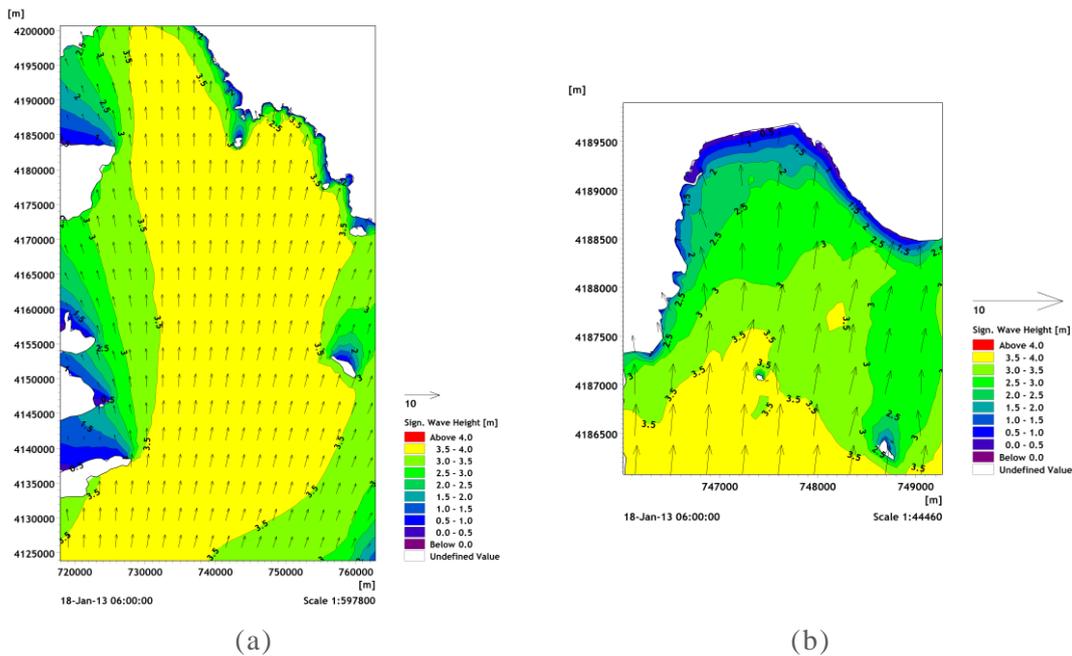
Figure 5-5. Spatial distribution of (a) significant wave height for the entire model domain and (b) for Varkiza bay at a specific time step of the simulation.

ST results

In Figure 5-6(a), the spatial distribution of the seabed level change at the specific time step is presented. Based on the simulation results, negative seabed level changes (up to -0.3 m) are observed along the coastline of the examined beach, while off the coast of Varkiza the corresponding seabed level changes are relatively smaller. Positive seabed level changes are depicted mainly along the east side of Varkiza Bay that may be attributed to the high values of current speed.

In Figure 5-6(b), the total load of sand transport is presented, along with the corresponding direction, for the examined extreme event. The highest values of sand transport (up to 0.00098 m^3/(s m)) are depicted mainly at the 4-m isobath at the central and eastern side of the beach, denoting erosion trend at a larger spatial scale compared to the west side. In the western part of the beach, at a zone of 150 m width from the coastline, accretion patterns are encountered while the rest zone is characterized by erosion. The same behaviour was revealed and discussed in the study of Skanavis (2013), where cross-shore profiles were obtained from a topographical survey by using RTK-GPS, and six sections (section A to F going from east to west, respectively) were presented along Varkiza beach before and after extreme events. In this work, three out of six cross-shore profiles, shown in Figure 5-7(a), are examined with reference to the period from January 5 to February 18, 2013.

In Figure 5-7, the seabed level change at the cross-shore sections (A, C, E) between the two examined dates (close to the beginning and end of the simulation period) is plotted, along with the initial section bathymetry. The changes calculated by the model are shown by using solid lines and the measured data by using symbols. It is revealed that at all the examined sections there is a clear erosional trend alongshore apart from the field measurements at section E, where accretion is observed for a distance approximately 15 m from the shoreline.
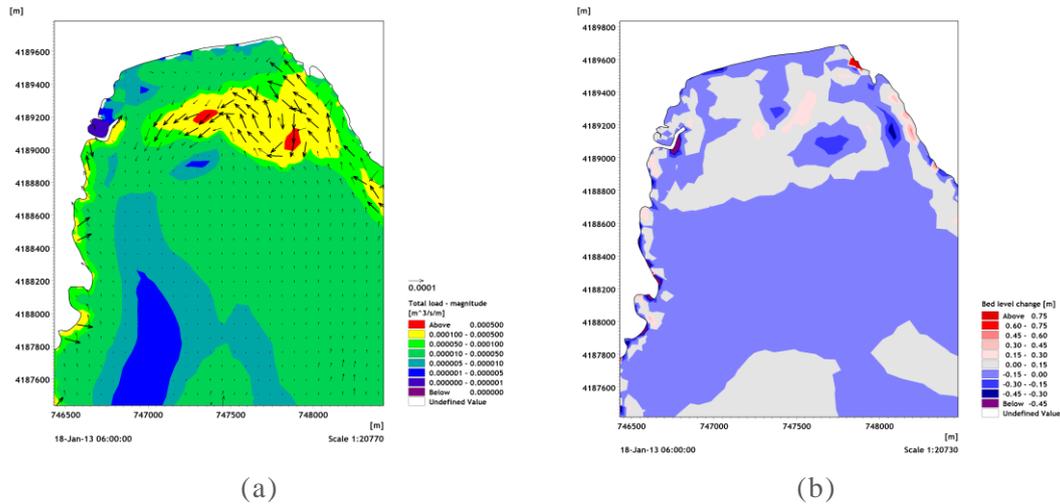
Figure 5-6. Spatial distribution of (a) seabed level change and (b) total magnitude of sediment transport for Varkiza bay at a specific time step of the simulation.

5.4.7 Discussion

Over the last 50 years, the human activities taken place in Varkiza, such as the construction of a high-traffic coastal avenue parallel to the beach at a distance about 180 m, and marine structures for mooring small boats and the needs of the local fishery community at the west side of the coast, have disturbed gradually the natural equilibrium between coastal hydrodynamics and sediment transport processes, and coastal configuration as well.

Based on the main findings of this study and the overview of the hydrodynamic conditions and wave climate of the beach, coastal protection measures and mitigation methods for coastal erosion at the examined area can be roughly suggested. As was stated by Bergillos et al. (2017), sustainable and economical interventions are preferred for coastal erosion problems; such countermeasures include, among others, beach nourishment (or beach fill), artificial reefs and coral transplantation known as soft engineering methods while breakwaters and other engineering structures belong to the hard engineering measures (Luo et al., 2016). The implementation of the former measures is also enhanced by the topography of Varkiza beach, since pocket beaches suffer less from lateral volumetric losses compared to open and extensive sandy beaches.

Whichever countermeasures will be adopted by the collaboration of coastal managers, scientists, decision makers and local authorities for the sustainable development and effective management of this coastal zone, previous extended video monitoring of the beach conditions, including periodically updated bathymetric data, is suggested. Furthermore, advanced local-scale shoreline evolution models, as e.g. UNIBEST (https://www.deltares.nl/en/software/unibest-cl/#8), requiring quite more detailed sedimentological information, allow for precise quantification of the sediment transport rates close to the shoreline.

(a)
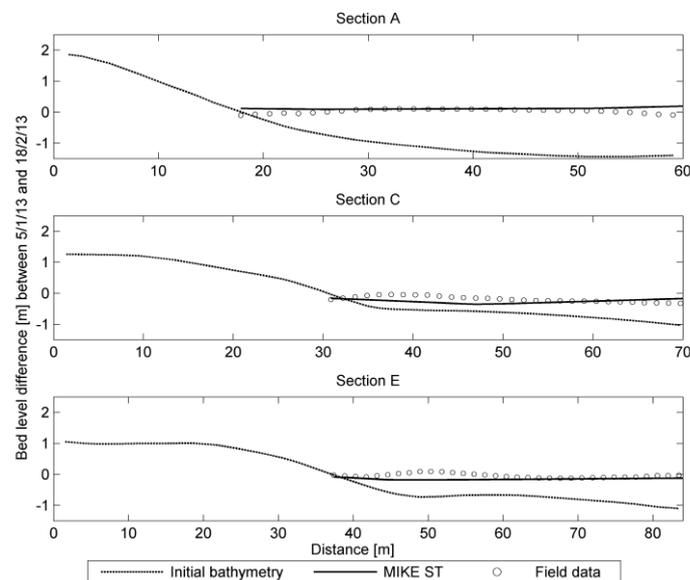
(b)

Figure 5-7. (a) Locations of cross-shore sections at Varkiza beach (from Google Earth) and (b) seabed level differences along the cross-shore sections A, C, E between January 5 and February 18, 2013 obtained from the model simulation and field data.

5.5   Case study 2: Sediment transport simulation based on the influence of cumulative wave action at a sandy beach

5.5.1   Motivation

In order to reduce computational time that is required for simulations of morphological models with time period of one year or greater, but retaining an acceptable accuracy of the predictions, wave input reduction methods have been suggested. The core idea of these techniques is to reduce the size of the wave input data at a coastal area of interest with some sets of representative wave conditions based on specific criteria.

In this section, a cost-effective method is introduced based on the use of process-based models combined with the philosophy of wave input reduction techniques. The proposed technique relies wave input reduction on a grain motion initiation criterion in terms of orbital velocity, from which two basic categories are separated: (i) the one dealing with wave conditions that contribute to the wave-induced initiation of sediment movement at depths around the closure depth, and (ii) the other one including the low energetic wave conditions. Other reference works as regards the onset of sediment motion under waves are those of Hallermeier (1980); Soulsby (1997); Van Rijn (1993). Consequently, the computational efficiency of estimating seabed level can be drastically increased with the proposed methodology instead of using the full wave time series, while the accuracy level can be retained into acceptable limits.

As an application the coast of Sitia, in the eastern part of Crete Isl., is examined as a specific case study. The main reasons for selecting this particular coast lie in its vulnerability to erosion phenomena and its touristic character. In Foteinis and Synolakis (2015), the mean coastal retreat rate at Sitia was estimated at 0.32 m/yr, among the highest erosion rates in Crete, utilizing aerial photographs (1960–2004), satellite images (2003–2012) and field survey measurements (2009–2012). In 2016, the collapse of the retaining wall of the coastal road brought the erosion matter to a climax leaving some villages in the north-eastern part inaccessible until the end of repair works. However, from Google Earth images, it seems that in 2017 there was a widening along the beach, which can be attributed to natural processes since no beach nourishment took place. Additionally, the touristic activities in the wider area have become more intensive the last years rendering confronting, prediction and management of erosion even more imperative. A preliminary study as regards the sediment transport patterns under two alternative wave scenarios (i.e. mean sea state, harsh wave conditions that contribute to initiation of sediment motion) and three different topographies of the seabed (i.e. current state, two submerged breakwaters at the isobaths of 5 m, port extension in the sea) has been conducted by the same authors at the same study area (Karathanasi et al., 2017). One of the main conclusions of this study as regards the harsh wave conditions for all the examined seabed topographies was the clockwise current circulation that contributed to the sediment movement westward.

The results from this analysis have been published in:
Karathanasi, F., Belibassakis, K., 2019. **A cost-effective method for estimating long-term effects of waves on beach erosion with application to Sitia bay, Crete**. *Oceanologia* 61(2): 276–290.

5.5.2   Methodology

When a long-term time series of wave data is available near-shore, the core of the proposed methodology is based on the rationale of wave input reduction. The wave conditions that contribute to the onset of sediment motion below the closure depth of a sandy seabed level, called hereafter "over-critical wave conditions", form the determinative factor of this analysis. With the term "closure depth" is defined the transition zone in which the influence of waves on bed stresses, and hence sediment transport, is significantly lower than within the region of wave breaking (i.e. surf zone) or the region where the effects of wave energy dissipation are dominant (i.e. upper shoreface zone) (Ortiz and Ashton, 2016). Hence the underlying assumption as regards closure depth is its dependence on the harsh wave conditions. In this context, it is possible to significantly reduce computation times and speed up the whole analysis. The proposed approach uses the wave statistical parameters such as significant wave height H_S and peak period T_P, along with some basic hydrodynamic parameters (e.g. wave height, sea water density) and sediment characteristics (e.g. d_{50}, density of sediment), to estimate bottom orbital velocity u_b and wave shear velocity u_{*w}, rendering the methodology fully applicable and handy, since in the majority of the cases such summary data are available (e.g. wave model outputs, archived wave data).

Case study 2: Sediment transport simulation based on the influence of cumulative wave action at a sandy beach

Before proceeding with the description of the methodology, for the sake of simplicity, let us first provide the appropriate definitions regarding the points used in the analysis that are mentioned in the subsequent sections:

- the offshore points that correspond to the available wave time series, forming the input for the boundary of the outer model domain with the coarse spatial resolution, are denoted by P_{out};
- the points that are used as input for the boundaries of the inner model domain with the fine spatial resolution, obtained after applying a wave transformation scheme, are denoted by P_{inn}, and the middle point of the northern boundary is denoted by $P_{inn,m}$;
- the point that represents the closure depth is denoted by P_{cd}, and the corresponding depth h_{cd} is defined by the Hallermeier (1981) equation given by:

$$h_{cd} = 2.28H_{eff} - 68.5\left(\frac{H_{eff}^2}{gT_{eff}^2}\right),$$

(5.6)

where H_{eff} is the effective wave height, exceeded 12 h in a single year (i.e. the greatest 0.137% waves during a year) and T_{eff} is the associated wave period.

Description of the cost-effective method

According to linear wave theory, the bottom (or near-bed) orbital velocity of a monochromatic wave is related to water depth and surface wave conditions as follows:

$$u_b = \frac{\pi H}{T\sinh(kh)},$$

(5.7)

where H is the wave height, T is the wave period and $k = 2\pi/\lambda$ is the wavenumber (λ is the wavelength) and h is the water depth. Eq. (5.7) is extended for multichromatic waves in the coastal environment by applying it for all frequencies of the wave spectrum corresponding to each sea state and summing the components. Thus, a representative bottom orbital velocity u_{br} is calculated; see, e.g. Madsen (1994). Following the method suggested by Wiberg and Sherwood (2008) a generic form of the wave spectrum is used to estimate bottom orbital velocity from the values of H_s and T_P of the reference wave data (i.e. the entire time series of the available wave data) at a point that represents the closure depth, denoted by P_{cd}. Among the commonly used wind-generated wave spectra, JONSWAP spectrum (Hasselmann et al., 1973) is adopted

$$S_\eta(\omega) = B\left(\frac{H_S}{4}\right)^2 \frac{\omega_P^4}{\omega^5}\exp\left[-\frac{5}{4}\left(\frac{\omega}{\omega_P}\right)^{-4}\right]\gamma^{\phi(\omega/\omega_P)},$$

(5.8)

where $\omega_P = 2\pi/T_P$ is the peak angular frequency, $B = 3.29$, $\gamma = 3.3$ and $\phi(\omega/\omega_P) = \exp[-0.5\beta^{-2}(\omega/\omega_P - 1)^2]$ with $\beta = 0.07$ for $\omega \le \omega_P$ and $\beta = 0.09$ for $\omega > \omega_P$.

The representative orbital velocity u_{br} is then calculated from the following relation

$$u_{br} = \sqrt{2\left(\sum_i S_{u,i}\Delta\omega_i\right)},$$

(5.9)

with $S_{u,i} = \frac{4\pi^2}{T_i^2\sinh^2(k_i h)}S_{\eta,i}$.

For the sediment transport purposes, another important property of waves is the bed shear stress τ_{bw} that can be associated with u_b and a wave friction factor f_w by Eq. (5.4). In this study, the friction factor is calculated by the following empirical relationship

$$f_w = \begin{cases} 0.04 \left(\dfrac{\alpha}{k_N}\right)^{-0.25}, & \dfrac{\alpha}{k_N} > 50 \\ 0.4 \left(\dfrac{\alpha}{k_N}\right)^{-0.75}, & \dfrac{\alpha}{k_N} < 50, \end{cases} \tag{5.10}$$

where $\alpha = 0.5H/\sinh(kh)$ is the wave orbital amplitude and k_N is the Nikuradse's bed roughness parameter equal to $2.5d_{50}$.

Wave shear velocity u_{*w} is defined as follows:

$$u_{*w} = \sqrt{\frac{\tau_{bw}}{\rho_w}}. \tag{5.11}$$

The dimensionless bed shear stress, i.e. the Shields parameter θ^*, defined as:

$$\theta^* = \frac{u_{*w}^2}{(s-1)g d_{50}}, \tag{5.12}$$

with $u_{*w}^2 = 0.5 f_w u_{bw}^2$ (where maximum orbital velocity is calculated using the significant wave height), $s = \rho_s/\rho_w$ denoting the ratio between the density of bed material and sea water (ρ_s is the density of the sediment) and g denoting the acceleration caused by gravity (9.81 m/s^2), is used to indicate the lower threshold value for initiation of sediment motion for the cases that $\theta^* > \theta_{cr}$, where $\theta_{cr} = 0.045$ isthe critical bed shear stress.

Based on the above threshold value of initiation of sediment movement, the proposed methodology can be applied on the available wave time series at P_{cd} in order to indicate the specific timesteps that represent these wave conditions yielding a value of θ^* higher than 0.045 (i.e. over-critical wave conditions). Let us note that in case the available wave time series is available at an offshore location, like P_{out} points, a wave transformation process should be necessarily implemented in order to obtain the corresponding time series at the closure depth. Having these over-critical wave conditions at P_{cd} to hand, the corresponding conditions at the boundary of the inner model need to be extracted, represented by $P_{inn,m}$. Since the temporal resolution of the wave time series is 1 hour and given the distance between the offshore boundary (of the inner model) and P_{cd} (~1.6 km), the over-critical wave conditions at the boundary of the inner model that contribute to the initiation of sediment motion are identified based on the same timestep that gives each over-critical wave condition at P_{cd}. Then, these over-critical conditions are classified at $P_{inn,m}$ into specific intervals of H_S and T_P (0.5 m and 1 s, respectively) with equidistant binning (i.e. constant bin-size) and the corresponding mean wave direction um is calculated for each class. This schematization (into (H_S, T_P, θ_m) triplets) is essential in order to proceed with the proposed methodology described in detail in the remaining part of this section.

Apart from the over-critical wave conditions, in which the morphological changes are large, the conditions where wave-induced currents are dominant should be additionally considered for a more realistic long-term behaviour of seabed level. Assuming that waves below 0.5 m at the boundary of the inner model do not produce significant erosion/accretion patterns in the shore, the calm wave climate, called hereafter "sub-critical wave conditions", is grossly classified for values of H_S smaller than the threshold values and higher than 0.5 m. In this case, the intervals

for H_S remain 0.5 m and for T_P the interval is varying (from 1 s to 4 s). The corresponding mean wave directions θ_m for the selected pairs (H_S, T_P) is also calculated.

The final triplets of both the over- and sub-critical wave conditions comprise the input for MIKE 21 Coupled Model Flexible Mesh (called hereafter MIKE21 CFM) simulations, which is the process-based model used in this work; see also Sections 5.5.4 and 5.5.5. From these simulations the rate of seabed level change q is extracted for a 2-week simulation period with 1-hour timestep. This time period allows a detailed sediment response for the specific triplets and a more accurate estimation of a mean rate q. Let us note that the rates estimated for the over- and sub-critical wave conditions are appropriately weighted based on the frequency of occurrence of each selected class.

After the schematization of the over- and sub-critical wave conditions, from the simulation results, the rate of seabed level change is estimated based on the sediment continuity equation. The mean rate of seabed level change q [m/day] for each triplet is calculated by

$$\bar{q} = \frac{\sum_{i=2}^{n} q_i}{n - 1}, \tag{5.13}$$

where n is the total number of timesteps during the 2-week simulation period. The rate of the first timestep q_1 is considered as an initialization rate of the simulations and for this reason, it is excluded from Eq. (5.13).

For the proposed methodology, the seabed level is estimated by

$$h(jt) = h(jt - 1) + \bar{q}, j = 1, \dots, n, \tag{5.14}$$

at the t −th 1-hour interval for each (H_S, T_P, θ_m) triplet.

Based on the above mentioned description and definitions, the frame of the cost-effective methodology is presented in Figure 5-8. Recapitulating the steps that should be followed for implementing the proposed methodology, the following key-aspects should be addressed:

1. Obtain wave time series at P_{inn} points and P_{cd}, if wave data are only available offshore;
2. Calculate bottom orbital velocity, wave shear velocity and bed shear stress at P_{cd};
3. If $\theta^* > \theta_{cr}$ at P_{cd}, then identify the corresponding values of H_S and T_P at P_{cd}. Based on the timestep of each pair, extract the corresponding over-critical values of (H_S, T_P) at $P_{inn.m}$. Then, group these pairs and calculate mean value of θ_m for each class;
4. If $\theta^* \leq \theta_{cr}$ at P_{cd}, then identify these values of H_S that are both higher than 0.5 m and different from the over-critical values (from step 3) along with the corresponding values of T_P. Then, group these pairs and calculate mean value of θ_m for each class;
5. Calculate the rates of seabed level change with MIKE21 CFM for both over- and sub-critical values for each (H_S, T_P, θ_m) triplet;
6. Finally, calculate seabed level at any location of the inner model domain via Eq. (5.14).

5.5.3 Case study

The area of interest is Sitia beach that is located in the north-eastern part of the Prefecture of Lassithi, Crete, on the west side of the homonymous bay; see Figure 5-9. It is a 2-km long beach with variable width of maximum value around 35 m, and exhibits a typical U-shape in the NW-SE orientation. Due to the shape and orientation of the examined beach, the wave action is confined to the north and north-eastern directions, which is the primary factor for the settlement of sediments. At the western part of the beach there is a river system (Pantelis -or Stomios-

river), following dry and wet periods, that discharges into the bay, and there is also the homonymous port that can accommodate both small fishing vessels and larger merchant and passenger vessels.
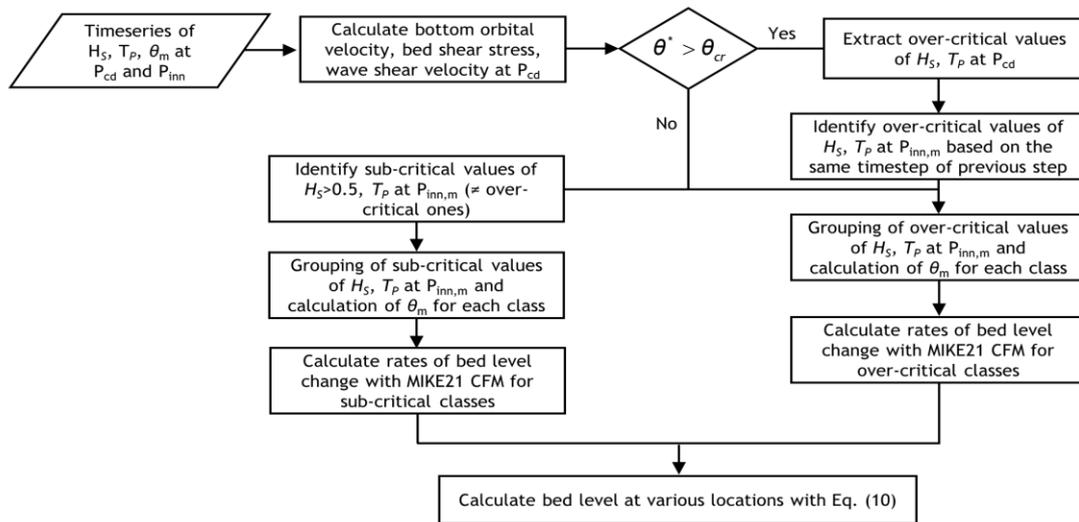
Figure 5-8. Flow chart of the proposed methodology.

Figure 5-9 also presents an overview of the points mentioned in Section 5.5.2 for the case study of this work. In this case study $h_{cd} = 6.5$ m, thus P_{cd} was selected on the isobath of 6.5 m and in the middle of the longshore direction of the beach.

The homonym town, Sitia, has become a tourist attraction the last decades, mainly during the summer period, while tourist infrastructures (e.g. hotels, restaurants), and in general, human activities, place pressure on the coastal environment. Moreover, the main road that connects Sitia with other tourist destinations at the eastern part of the island, such as the palm forest Vai, was developed to a great extent beside the coastal front.

To this end, erosion phenomena are evident due to both the intensive residential and infrastructure-based development of the wider area along with the physical conditions that seem to be more frequent and of longer duration. Specifically, at the end of 2016 the front of the coastal road that is contiguous to the eastern part of the beach collapsed after the accumulative action of intense weather conditions that took place the last few years, causing several problems and safety issues to the local residents and tourists. Moreover, the sediment supply of the beach is relatively limited while the construction of the adjacent harbour at the western part of the coast, in order to serve the needs for tourism and fishing, puts additional pressures and intensifies erosion rates.
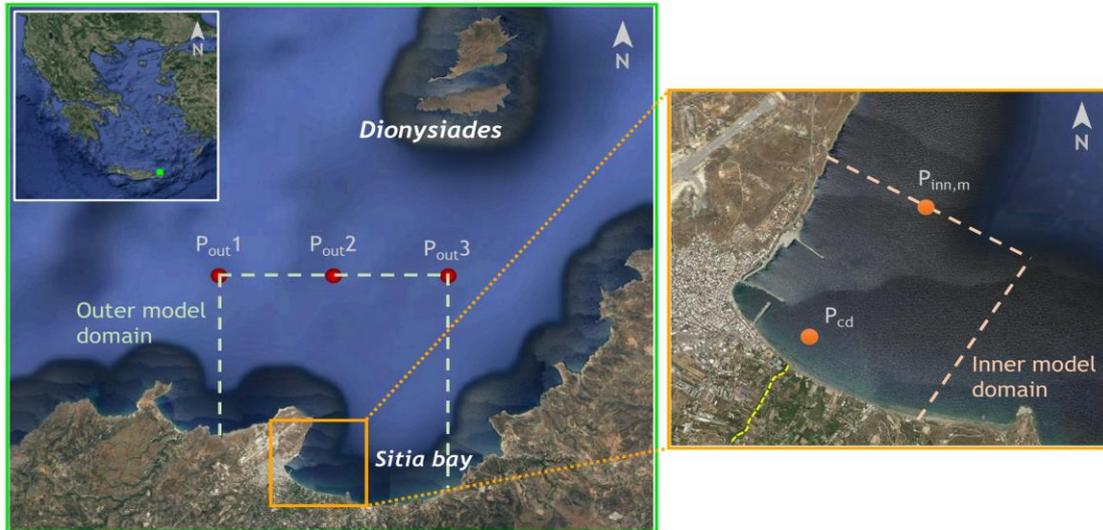
Figure 5-9. Aerial map of Sitia bay along with the offshore locations of the input data for the outer model domain (left map), and the study area of Sitia beach (inner model domain) along with the locations of $P_{inn,m}$ and P_{cd} (right map) used in the analysis. (Source: Google Earth)

5.5.4   Model setup

As mentioned above, the process-based numerical model that is used in this study is MIKE21 CFM developed by the Danish Hydraulic Institute (DHI). MIKE21 CFM is a depth-averaged two-dimensional numerical model used to study and simulate a wide range of coastal hydrodynamic problems including the description and interaction of the relevant processes, such as currents, waves and sediment transport in coastal areas, among others. This numerical modelling software package includes several interrelated modules, of which the following are used for the purpose of this study: (i) the hydrodynamic (HD) module; (ii) the spectral wave (SW) module, and; (iii) the sand transport (ST) module. Through a dynamic coupling, hydrodynamic and spectral wave computations are performed simultaneously to calculate sediment transport rates and update bathymetry at each timestep. Specifically, sediment modelling is established on: (i) a depth-averaged hydrodynamic model, based on the depth-integrated incompressible Reynolds averaged Navier-Stokes equations; (ii) a phase-averaged wave model, based on the wave action conservation equation, and; (iii) sediment transport tables calculated in advance for every combination of current, wave, bathymetry and sediment conditions appearing in the simulation; for a more detailed description of the three modules, see Belibassakis and Karathanasi (2017).

In the following subsections, the boundary conditions and the model parameters used for the model simulations are described for each module, along with some necessary information as regards the model grid and wave climate.

Bathymetry and unstructured grid

As already mentioned, in this analysis, the outer model domain is used for the transformation of the wave conditions from the available wave time series towards the shore. This model domain covers a distance of 7.5 km in the longshore direction and 7.8 km for the cross-shore one. The total number of triangular elements in the outer domain is 1,284 with 759 nodes while

the maximum size of the elements is approximately 0.12 km^2; see also Figure 5-10(a). The bathymetry of the outer model domain presented in Figure 5-10(b), shows that the seabed topography is quite mild. From the shoreline up to the isobath of -75 m, the contours are parallel and the maximum depth (-226 m) is observed at the north-western part of the domain.
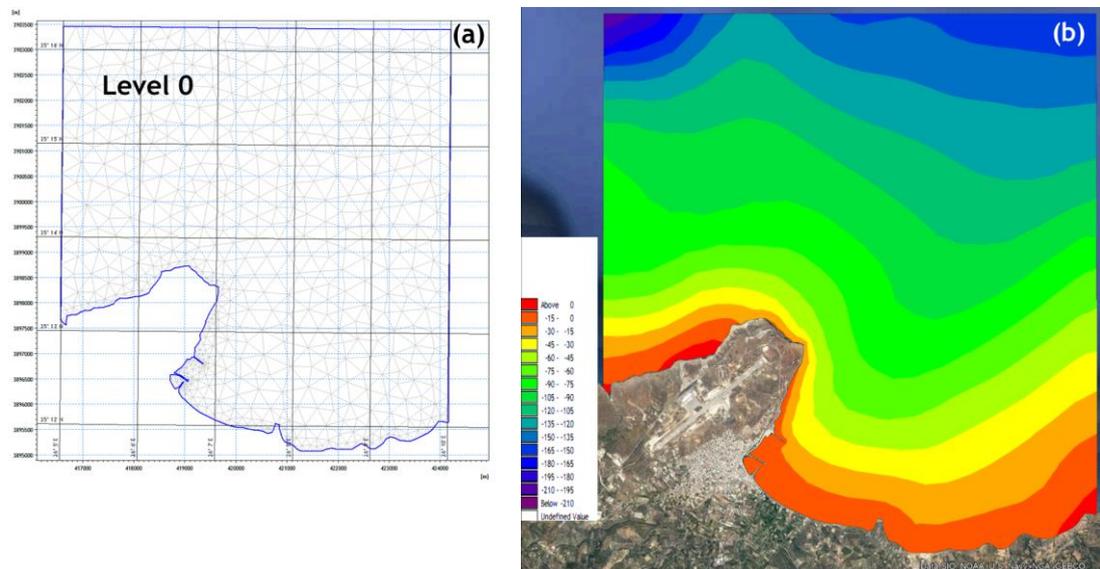
Figure 5-10. (a) Mesh with triangles for the outer model. (b) The bathymetry of the outer model domain.

As regards the inner model domain, it is divided into two nested grid domains, going gradually from the outer area with the lower resolution (i.e. level 1) up to the computational grid with the highest resolution (i.e. level 2), where the smaller triangular elements represent areas where the accuracy in the wave, current and sediment transport calculations are important; see also Figure 5-11(a) for the representation of the different levels and the final mesh generation of the examined area. Specifically, level 1 extends both in the longshore and cross-shore directions approximately 1.7 km with the area of each triangular element not exceeding 6,580 m^2. Let us note in advance that the appropriate forces are imposed at the boundaries of the outmost level (i.e. level 1) for the generation of flow and wave conditions, which in turn define the corresponding boundary conditions of the inmost level (i.e. level 2). The second, and more detailed, computational grid (level 2) extends in the longshore and cross-shore directions 1,400 m and 140 m, respectively, with maximum area of each triangular cell up to 1,050 m^2. The total number of grid cells in the inner domain is 2,135 with 1,282 nodes.

The bathymetry data of the inner model domain were digitized from maps of different spatial scales obtained from the Hellenic Navy Hydrographic Service (HNHS). The above data were enriched for the outer model domain with bathymetric grid points from the European Marine Observation and Data Network (EMODnet) Digital Bathymetry database with 1/8 of an arc minute (~230 m) resolution (Marine Information Service, 2016).

In Figure 5-11(b), the 2D bathymetric representation of the study area is displayed in Google Earth for levels 1 and 2. The isobaths from -20 m to lower depths are generally parallel to the shoreline and are evenly flattened going from the offshore part towards the shore. The highest depth (close to -50 m) is encountered in the north-western part of level 1 while the 10-m isobath is about 410 m from the coastline. In the eastern part of Sitia beach, there are beachrocks aligned parallel to the shoreline starting approximately from -1.5 m depth and ending to the coast. The formations act as natural submerged breakwaters mitigating erosion phenomena at this part of the coast.
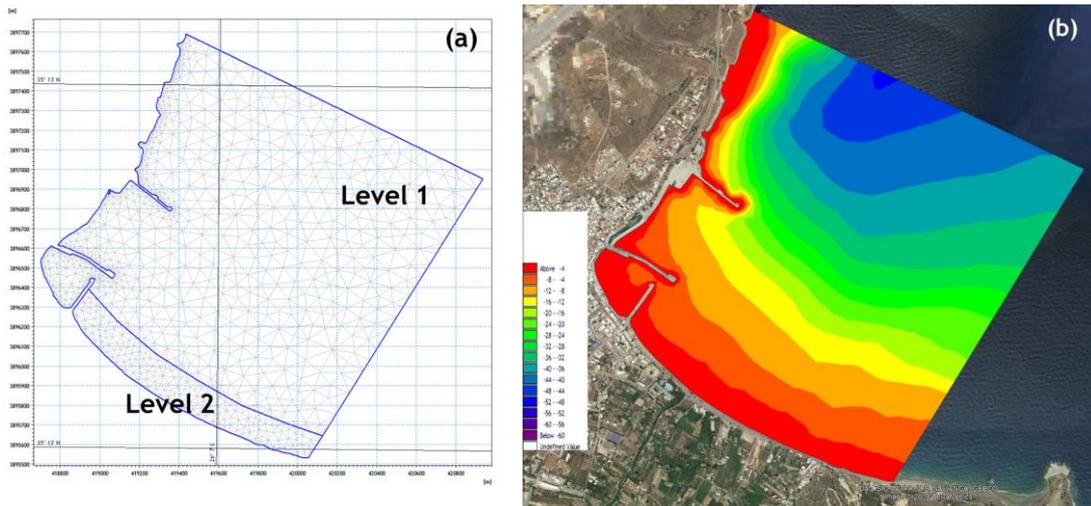
Figure 5-11. (a) Mesh with triangles using two levels for the inner model domain. (b) The inner model domain showing the bathymetry of the examined area.

Wave conditions

As regards the wave characteristics of the wider study area, the analysis relies on 1-year time series, between 01/01/2016 and 31/12/2016, at the middle point of the boundary of the outer model, i.e. at $P_{out}2$ (see also Figure 5-9, left map), with geographical coordinates 35.271°N–26.125°E, obtained from the Mediterranean Sea Waves database. The relevant information include significant wave height H_S, peak wave period T_P and mean wave direction θ_m (measured clockwise from north), with an 1-hour resolution. These time series were used as input for the wave propagation from the offshore to the near-shore using MIKE21 CFM (SW and HD modules). After this simulation, the spectral time series were extracted for the northern and eastern boundaries of the inner model domain (with the finest triangular elements), presented in the right map of Figure 5-9, in order to be used as input for the rest simulations.

The basic statistical measures at $P_{out}2$ include mean value (*m*), standard deviation (*sd*), minimum (*min*) and maximum (*max*) values, 50th percentile (*p50*), skewness (*sk*) and kurtosis (*ku*), and the results are presented in Table 5-1. On average, the wave intensity is characterized low with mean values $m_{H_S} = 0.9$ m, $m_{T_P} = 5.08$ s and $m_{\theta_m} = 394.1°$. The most intense wave incident occurred on 6th February, 2016 with $H_S = 4.8$ m and corresponding $T_P = 9.23$ s and $\theta_m = 344.4°$ during a two-day storm. The value of sk_{T_P} (0.23), close to zero, indicates that the distribution of the corresponding data is close to be symmetrical while the highest value of *ku* (7.1) is given by H_S indicating a sharp peak of the distribution.

Table 5-1. Basic statistics of the wave parameters obtained from the spectral time series at $P_{out}2$ between 01/2016 and 12/2016. Square brackets denote units of the corresponding wave parameter where necessary.

	N	m	sd	min	p50	max	sk	ku
H_S (m)	8784	0.9	0.7	0.1	0.7	4.8	1.8 (–)	7.1 (–)
T_P (s)		5.08	1.53	1.37	5.21	10.15	0.23 (–)	2.9 (–)
θ_m (°)		394.1	0.6 (–)	–	396.7	–	−0.01 (–)	0.7 (–)
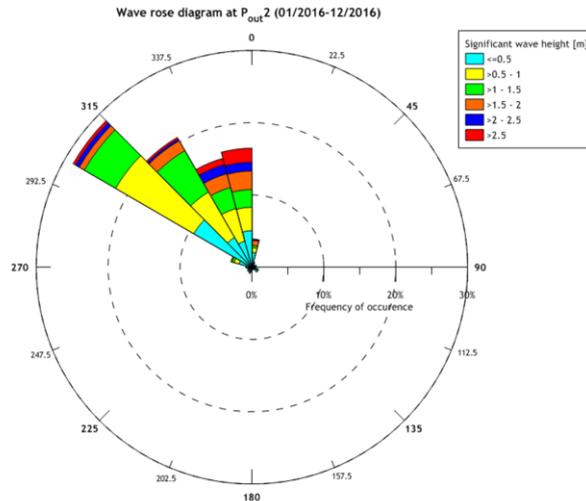
Figure 5-12. Rose diagram of significant wave height and wave direction at P$_{out}$2 for the period 01/2016–12/2016. Intervals for H_S and θ_m are $\Delta H_S = 0.5$ and $\Delta\theta_m = 15°$, respectively.

As regards θ_m, the low value of *sd* (0.6) corresponds to a circular dataset that is highly concentrated, which can be also verified in Figure 5-12, while *sk* value close to zero (−0.01) denotes a unimodal distribution. The wave rose of H_S at P$_{out}$2 is depicted in Figure 5-12, along with the corresponding frequencies of occurrence. The scattering of wave directions is limited to the sector due to the topography and coast orientation of the study area with the prevailing wave directions coming from the north direction (sector $[300°, 315°]$), which are attributed to the very large fetch (390 km). The highest frequency of occurrence (13%) as regards wave propagation in the dominant direction is observed for values of H_S between 0.5 m and 1 m while the corresponding values of T_P exhibiting the highest frequency of occurrence are between 4 s and 6 s. Intense sea states ($H_S > 2.5$ m) with the highest frequency of occurrence (2%) correspond to the sector $[345°, 0°]$.

Input data

For practical reasons, the period of the simulation is confined to one year, i.e. from January 1 to December 31, 2016. As already mentioned the bathymetry resolution for the inner model domain gets progressively finer as we move from level 1 to level 2, which is the area of interest as concerns the simulation results and the evaluation of the methodology. The timestep is set to $\Delta t = 3600$ s, equal with the time interval of the available time series. Prior to the description of the input data for the one-year wave time series, let it be mentioned that the authors kept some parameters at their default values since no *in situ* measurements were available for calibration of the model.

As regards HD module, the most essential input data include: wave radiation stress gradients that force the flows, bed resistance, eddy viscosity and boundary conditions. Eddy viscosity is based on the Smagorinsky coefficient with a constant value at 0.28, bed resistance expressed through the Manning number was fixed (32 m$^{1/3}$/s) in the entire inner model domain apart from its south-eastern part due to the presence of bedrock formations while density is not updated during the simulation (barotropic mode). Note that tidal potential is very low in Sitia bay thus it is not considered in the model setup. At the open boundaries, current velocities (varying in time and along boundary) are used as input obtained from the simulation results of the outer model while at the closed boundary, the normal velocity component is set to zero, assuming full slip boundary conditions.

As in the HD module, the instationary mode as regards time formulation was adopted as well in the SW module as well, with a directionally decoupled parametric formulation. The conditions at the open boundaries (at the north and east side of the model domain) were kept constant in space (along the boundary line) and varying in time while the boundary data consisted of significant wave height H_S, peak wave period T_P, mean wave direction θ_m and directional spreading index n_{ds}. Additional model parameters were wave breaking specified by the gamma parameter $\gamma_{wb} = 0.8$ constant in space, bottom friction specified by the Nikurdase roughness k_N, which was varying in space ranging from 6.25 mm to 0.25 mm for level 1, and 1.9 mm for level 2 while for the bedrock formations the value of 62.5 mm was selected.

Regarding the setting up of the ST module, sediment transport rates and seabed level changes under the combined action of waves and currents are calculated through interpolation of sediment transport tables. These tables are generated in advance and include the following parameters: root-mean square wave height, peak period, current speed, wave height-to-water depth ratio, angle between current and waves, median grain diameter d_{50} and sediment grading. The ST calculations are activated at the initial timestep while the timestep factor is set to 1, meaning that sediment transport rates and seabed level are calculated every timestep. Apart from the flow (HD) and wave (SW) forcings, the specification of sediment properties and the considerations of morphological impact on hydrodynamics are two important features that need to be provided for the area of interest. To this end, as regards the granulometric composition of the bottom sediments in the study area, the sea bottom consists of sand with an average diameter of d_{50} around 0.65–0.85 mm up to the isodepths of 1.5–2 m and with d_{50} between 0.08 mm and 0.25 mm for depths above 15 m (Anagnostou et al., 2017). Sediment grading was kept fixed, equal to 1.45, at the entire model domain. The initial bed layer thickness for all levels was set to 0.5 m apart from the bedrock part (0.0001 m).

In terms of the representative wave conditions (both over- and sub-critical ones), the parameters of the model setup remained the same except for the time formulation (quasi stationary mode) and the start time of the ST calculations since all modules were synchronized to start at the same timestep.

5.5.5   Results

Representative wave conditions

In this study, the time period of the analysed wave data is confined between 01/2016 and 12/2016; henceforth, when we refer to the full time series of 2016 we use the term "reference wave data". The time series of the reference wave data for H_S and T_P at $P_{inn,m}$ is presented in Figure 5-13. Consecutive intense wave conditions with $H_S > 1.5$ m occurred mainly during the last two months of the examined year. In the majority of the timesteps, high values of H_S correspond to high values of T_P as regards the examined location, rendering these pairs candidates for the initiation motion of sediments. According to the methodology, the first step is to calculate representative orbital velocity, bed shear stress and wave shear velocity by using the H_S and T_P time series of P_{cd} by applying Eqs. (5.4), (5.9) and (5.12), respectively. Based on the calculation of the Shields parameter and its threshold value, the over-critical wave conditions at P_{cd} are determined. Classifying the reference wave data at P_{cd} into classes of H_S and T_P with intervals 0.5 m and 1 s, respectively, we obtain Figure 5-14(b). From this figure it can be noticed that the lower threshold values for the onset of sediment transport, based on the Shields criterion, correspond to waves higher than 1 m with peak period between 6 s and 10 s and mean wave direction around 25°–29° as regards P_{cd}.
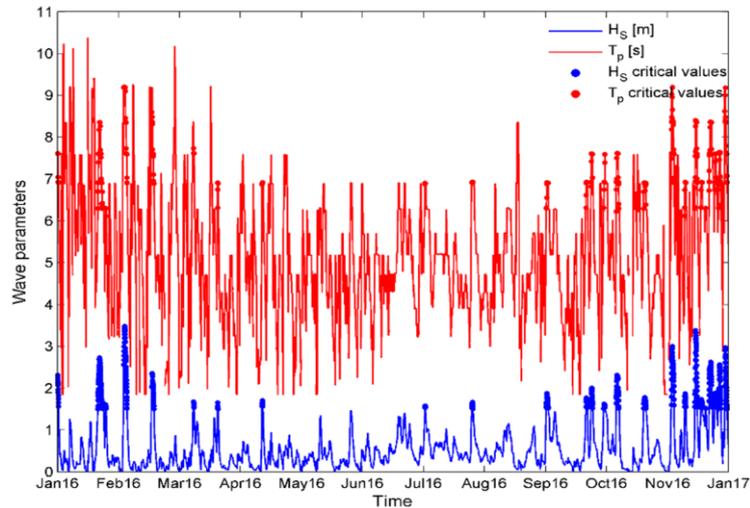
Figure 5-13. Time series of H_S and T_P at $Pi_{inn,m}$ for the year 2016. Blue and red dots indicate over-critical values of H_S and T_P, respectively.

Identifying the corresponding wave conditions at the boundary of the inner model, i.e. at $P_{inn,m}$, the corresponding threshold values are presented in Figure 5-14(a) with the blue outline having minimum values 1.5 m and 6 s for H_S and T_P, respectively, and in the range $[355°, 5°]$ for θ_m. As a whole, nine representative intense wave conditions (i.e. over-critical pairs) were taken into account for the estimation of rates of seabed level chanrg over the examined period. From the same figure, the calm (sub-critical) wave conditions were derived by further grouping these classes into eight representative calm wave conditions with the same interval for H_S and a varying one for T_P, depending on the bivariate histogram. Let us remind that small values of H_S (i.e. <0.5 m) are not considered in the next steps of the technique since the model runs of the sensitivity analysis, performed by the same authors, demonstrated that such waves present almost negligible quantities of sediment transport rates. Altogether, 17 (H_S, T_P) pairs, along with the corresponding values of θ_m, are considered in the analysis, which were simulated separately.
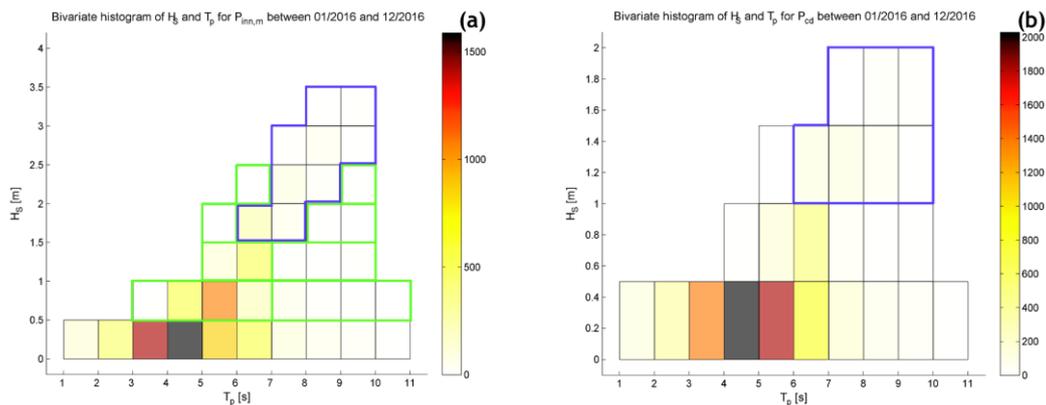
Figure 5-14. Bivariate histogram of (H_S, T_P) for (a) $P_{inn,m}$, and (b) P_{cd} for the year 2016. The blue closed lines indicate the over-critical values and the green rectangles indicate the sub-critical pairs.

Application of the methodology at the examined coast

Eight shallow locations are selected for examining the methodology described in Section 5.5.2; their geographical location, depth and distance from shore are given in Table 5-2. These points cover a distance of approximately 1,100 m along the coast with their in-between distance being around 150 m; their location on the map is shown in Figure 5-15(a).

Table 5-2. Name of location, geographical coordinates, depth and distance from shore.

Location	Geographical coordinates (lon, lat) (°)	Depth (m)	Distance from shore (m)
A	(26.1090°, 35.2060°)	-1.23	26
B	(26.1101°, 35.2050°)	-1.42	38
C	(26.1113°, 35.2041°)	-1.38	37
D	(26.1129°, 35.2030°)	-1.08	37
E	(26.1143°, 35.2024°)	-1.57	41
F	(26.1158°, 35.2017°)	-1.02	39
G	(26.1172°, 35.2013°)	-0.87	40
H	(26.1188°, 35.2007°)	-0.58	45
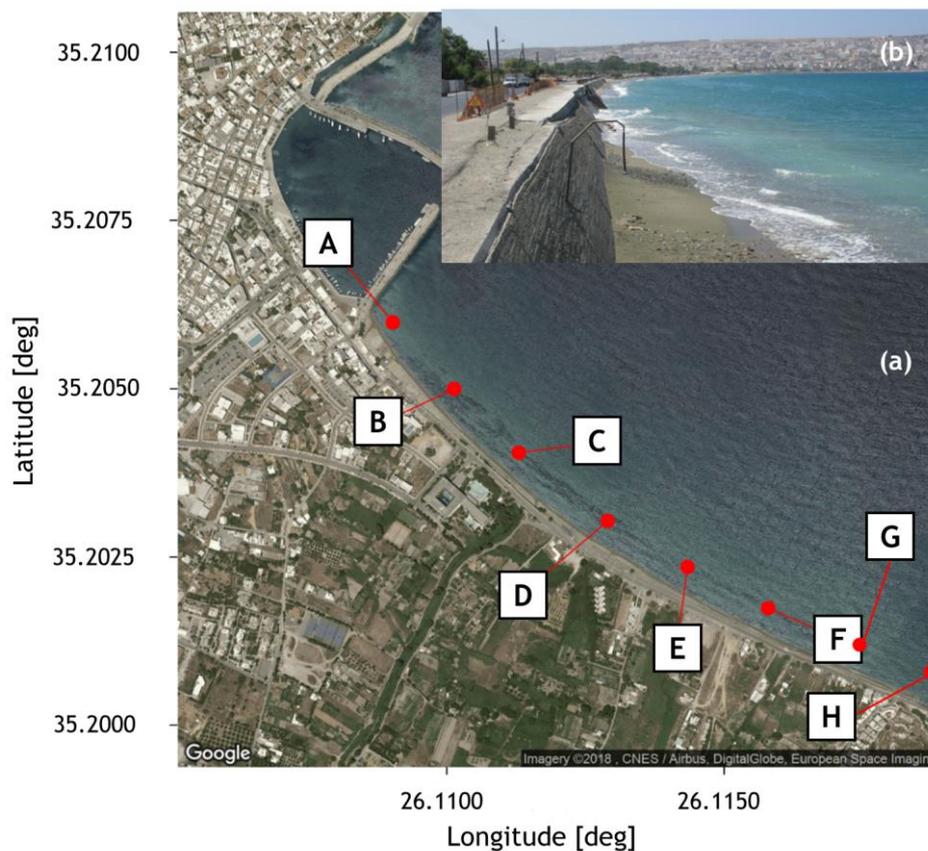
Figure 5-15. (a) Map of the examined area (from Google Earth) indicating the locations for the estimation of seabed level based on the proposed methodology at Sitia beach. (b) Photo near location G indicating erosion problem.

As regards the over-critical (H_S, T_P) pairs, the values of rates of seabed level change for locations C, D, E, F, G and H are negative, with values between −0.003 m/day and −0.003 m/day and −0.036 m/day. In general, the eastern locations (i.e. E, F, G and H) present the

highest negative rates of seabed level change while the western locations A and B are characterized by negative and positive rates of varying magnitude. With respect to the sub-critical (H_S, T_P) pairs, smaller, negative and positive, rates of seabed level change are provided by all locations compared to the above pairs with the highest positive value (0.034 m/day) encountered at location E and the highest negative value (-0.033 m/day) at location F.

A more analytic representation for estimating seabed level with the proposed methodology is given in Figure 5-16 for location A, and in Figure 5-17 for location F regarding specific representative (H_S, T_P) pairs. In the left panels of the above figures, the vertical lines denote the time windows of the over-critical (H_S, T_P) pairs in terms of sediment initiation; in the examined annual time scale, 30 time frames were identified by the methodology. In the right panels of the same figures, the rates of seabed level change are plotted for the two different types of representative wave conditions (i.e. over- and sub-critical). As it was expected, the rates of seabed level change for the over-critical (H_S, T_P) pair present higher values compared with the sub-critical pairs at both locations.
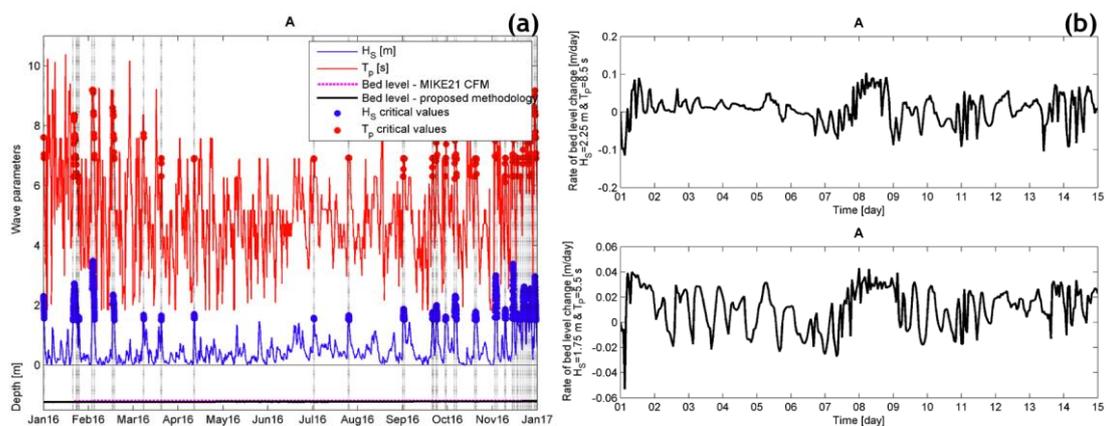
Figure 5-16. (a) Wave parameters along with seabed levels obtained from the two approaches. (b) Rates of seabed level change obtained from the proposed methodology for one over-critical and one sub-critical representative wave condition at point A
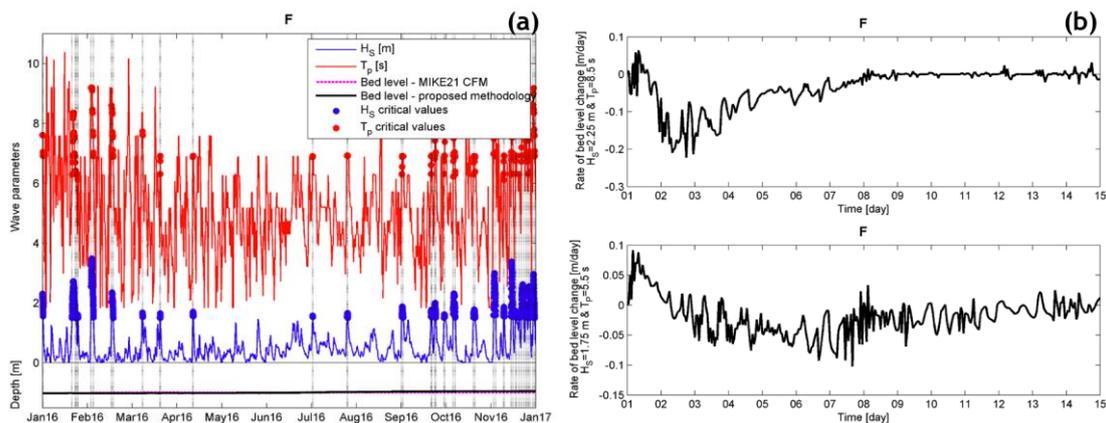
Figure 5-17. (a) Wave parameters along with seabed levels obtained from the two approaches. (b) Rates of seabed level change obtained from the proposed methodology for one over-critical and one sub-critical representative wave condition at point F.

Specifically, as regards location A, the pattern of the over-critical wave case shows some resemblance with the sub-critical one; in both cases, the rate of seabed level change strongly fluctuates during the 2-week simulation, taking mainly positive values, while at the 8th day of simulation a relative stabilization is evident. On the other hand, for location F, the rates present a dissimilar behaviour from location A; the rate of seabed level change seems to be stabilized around zero after eight days of simulation for the over-critical representative wave conditions while it takes constantly negative values, after the second day of simulation, with bigger fluctuations for the sub-critical ones.

In Figure 5-18, the values of seabed level obtained from the simulations results of MIKE21 CFM with the reference wave data as input, represented by the dashed line, and the proposed methodology, represented by the solid line, are plotted at the examined locations. From this figure, the following comments can be summarized:

- Locations B and D exhibit a very good agreement between the two approaches; throughout the year, the corresponding seabed levels follow the same tendency and are very close with each other while as regards the last month, the deviation between the two seabed level values is 0.7 cm and 0.4 cm, respectively, which are the smallest differences among the examined cases.
- Locations A and C, which follow a bathymetric profile with smooth to intermediate slopes (not shown here), and location H as well, exhibit medium-size deviations at the end of 2016, with values between 2.2 cm and 2.8 cm, respectively; however, the resemblance of the pattern that the two lines follow throughout the year is rather poor.
- The locations E and F, with the latter having a steep bathymetric profile, exhibit the second largest deviation at the end of 2016 (4.5 cm) but the lines indicating the seabed levels are in accordance in terms of the trend.
- Location G (see also Figure 5-15(b)) presents the highest deviation (6.6 cm) compared with the reference time series.
- The seabed level slope at locations A, B and C is positive indicating accretion in the western side of the Sitia coast while location D is characterized by a small negative slope (i.e. erosion pattern). Locations E and F present a steeper positive slope than the western locations, and locations G and H exhibit a higher negative slope than location D, implying more distinct erosion patterns. Overall, this behaviour coincides quite satisfactorily with the real situation encountered in the Sitia coast during the examined period, where the eastern part has been eroded to a great extent leading to the collapse of the retaining wall of the coastal road.

5.5.6   Discussion

The scope of this work was to reduce the reference wave data (of one-year duration) into two groups, i.e. (i) the over-critical (H_S, T_P) pairs that fulfil the Shields criterion leading to sediment initiation, and (ii) the sub-critical (H_S, T_P) pairs that do not fulfil this criterion, in order to significantly reduce computational times and compare the estimated seabed level values with the full case. The results of the proposed methodology compared to the ones obtained from utilizing the entire time series of the available wave data present similar trends, and the differences remain under 7%.

In this connection, some notable aspects should be remarked. Various sources of uncertainties as regards the discrepancies can be attributed to the assumptions that are imposed throughout the adopted technique. For instance, turbulence caused by wave breaking is not considered although it can be a source of sediment mobilization. Other uncertainties deal with the calculation of bottom orbital velocity, related indirectly with the Shields criterion, that does not take into account the presence of currents while the assumed spectral form might also influence bottom orbital velocity. For more details in terms of the potential sources of error in the

calculation of bottom orbital velocity from wave spectral parameters such as H_S and T_P, see further assumptions provided by Wiberg and Sherwood (2008). Furthermore, in the context of the sensitivity analysis, the authors followed an alternative way to estimate bottom orbital velocity and friction factor. The corresponding values derived from the simulation results of the reference wave data reached common over-critical combinations of H_S and T_P.

Another potential source of uncertainty could be the estimation method of the mean rates of seabed level change. Many dissimilar ways were tested by the authors including mean rates from one week, different mean rates based on the (H_S, T_P) pairs and the examined location, mean rates calculated with a smaller time interval during the simulation runs etc. However, the adopted approach showed consistently better performance in terms of seabed level prediction.

Let us also highlight that a more proper and fair comparison would be to assess both results from model simulations with *in situ* measurements of seabed level at the site of interest. The absence of real measurements has a twofold effect: i) it places the comparison into relative terms, and ii) it renders model calibration infeasible, thus the model results *per se* should be used with caution. Nevertheless, such comparison is beyond the scopes of this study. Moreover, due to the lack of real measurements, it is also recommended not to apply speed-up techniques since they require careful calibration and validation.

Another worth-mentioning fact refers to the distribution of wave direction. Specifically, the window of wave directions that affect significantly the morphological (bed level) conditions of the examined beach is very narrow since in the majority northern wave directions are dominant. This feature along with the gentle bottom slope and the uniformity of the coast as regards its shape render the study area a simple and easy example to implement this methodology compared to more complex cases.
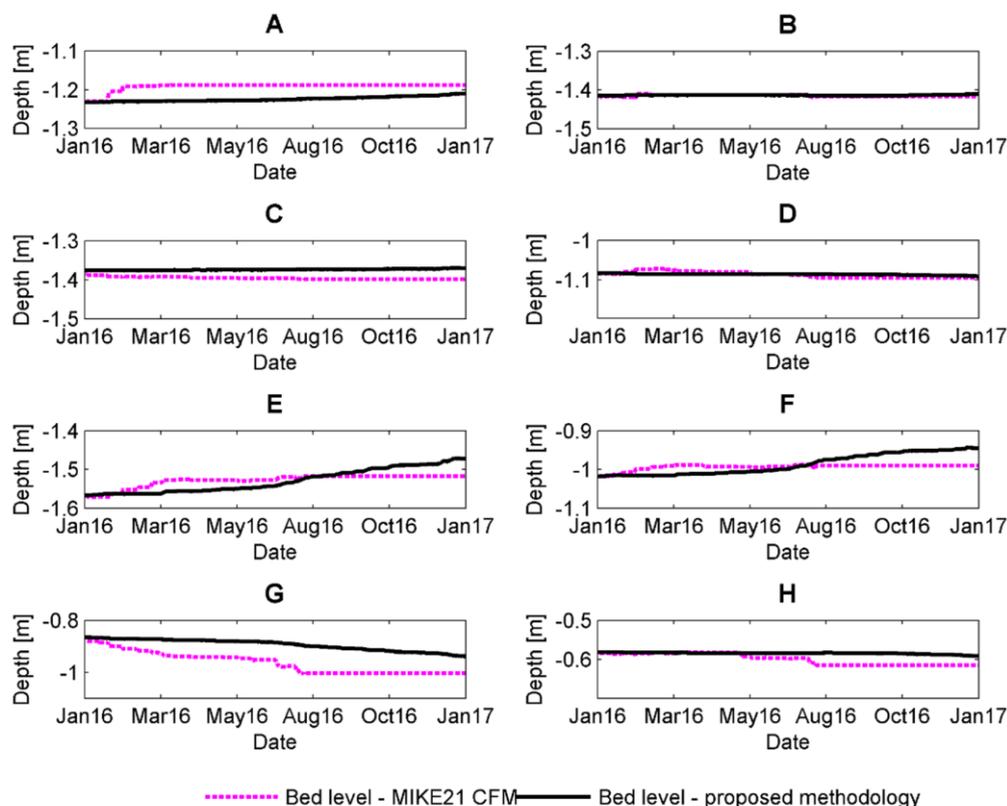
Figure 5-18. Seabed levels derived from MIKE21 CFM (dashed red line) and the proposed methodology (solid black line).

In reference with the overall computation time of the model simulations, there is a striking discrepancy between the two approaches. For MIKE21 CFM with the full reference wave data as input, the total runtime was 542 h while for the 17 representative wave cases of the proposed methodology, the corresponding runtime was 2 h. All simulations were conducted on an i7-2600 CPU server with 16 GB RAM and 3.40 GHz processor. Although current version of DHI is designed for parallel computing using graphics processors and could significantly accelerate the calculation process, still the present approach contributes to a significant runtime reduction, which for the particular non-parallel computing setup used is of the order of 99.6%. The latter result is quite impressive compared to the outcome presented in Figure 5-18, at least for the case-study examined, characterised by mild bottom topography and coastal characteristics, and regularly in the distribution of offshore wave directions.

Chapter 6 Conclusions and recommendations for future research directions

6.1 Summary

The objective that was at the core of this thesis was the probabilistic modelling of wind and wave variables, both linear and directional, with focus on the entire data sample, the extreme data and the relationship between two linear or directional variables. In this context, two specific areas of application that receive a lot of attention from the scientific community were selected for further investigation: i) the marine renewable energy sector, and; ii) the issue of coastal erosion. The two most promising forms of marine renewable energy for development in the Mediterranean Sea is offshore wind, with floating foundations offering the most optimum solution due to the deep continental shelf of this basin if the technological challenges are surpassed, and wave energy with numerous concepts and configurations but still with no reliable and cost-efficient solution towards commercialization. On the other hand, wind and wave forces are among the most important factors that affect erosion rates of a sandy beach. Apparently, these two research fields are also interconnected in many aspects; for instance, the presence of a wind (or wave) farm close to the coastal zone have impacts on the wave field and in turn, on the morphodynamics of the nearby coastal area due to the interactions between the turbines (devices) themselves, and with the adjacent coastal environment as well. Below, the main conclusions and contributions from this study are summarized.

Metocean climate variability has received a lot of attention in the relevant literature as the lifetime of marine renewable energy projects is planned for a long period (e.g. usually 30 years for an offshore wind farm). The main concern is energy generation, which is directly linked to various metocean variables. Since the spatial distribution and quality of wind/wave resource can be altered rigorously because of the climate change, the knowledge of climate variability can shed light on many aspects, such as planning phase and operating conditions. In this connection, when considering sufficiently long-duration metocean data, it was shown that long-term variations of metocean climate of both linear and directional characteristics should be identified and quantified not only for the annual (inter-annual) time scale but for seasonal and decadal (inter-decadal) cycles as well, indicating additional features and trends that might be associated with climate change. Moreover, diverse probabilistic models for linear and directional variables were quantitatively assessed in a consistent manner for both the univariate and bivariate case, revealing interesting aspects for the linear-directional modelling.

In offshore and coastal engineering applications, it is essential to work with metocean data of high quality. Among the main sources of uncertainty is uncertainties associated with the deficiencies that characterize each data source. Such uncertainties can be reduced by calibrating linear and directional data from the less accurate data sources, i.e. numerical models and satellite observations, with *in situ* measurements used as a reference source. Regression calibration method was used for this purpose with emphasis on robust estimators for linear variables, known to be efficient in the presence of outliers or when there are small deviations from the model assumptions. Statistical measures, such as bias, root mean square error and scatter index, were calculated from concurrent data to quantify the performance of the corresponding regression model and in turn, the relevant uncertainties. Calibration of directional variables was also performed in the present thesis with the results verifying its significance in engineering applications.

As already mentioned, the accurate analysis of extreme weather conditions plays a decisive role in marine renewable energy and provides basic information for the research on the design and in turn, reliability of the structure to withstand all environmental loads that is expected to face.

The directional description of the extreme climate of linear metocean variables is also important especially in planning and siting of a marine renewable energy project. On these grounds, the dependence of linear metocean variables with directionality was examined for performing extreme value analysis. Shape and scale parameters of the Generalized Pareto distribution were considered as functions of direction to account for this dependence in the estimation of design values, which was expressed by a Fourier series expansion due to its periodicity. Different methods for threshold selection and declustering were investigated and a penalised likelihood criterion was proposed for the estimation of the model parameters.

With regard to the problem of coastal erosion, the wave regime is a determinative factor in the formation of coastal shoreline and morphology. To this end, the response of sandy beaches to wave action under extreme (e.g. after a storm) and typical (e.g. considering the wave action during a year) conditions was examined by utilizing two different concepts, episodic and long-term erosion, respectively. In the former case, coastal erosion is mainly related to the predominance of high waves with a time span from some hours to days while in the latter one, the coast progressively adapts its form to the frequency and intensity of extreme events along with the prolonged calm wave conditions. Modelling of wave propagation, circulation and sediment transport was performed with the MIKE software. Specific cross-shore sections were selected to study the beach profile changes as regards episodic events at a beach, where measurements, available from a topographical survey were used as a reference source for comparison purposes. A cost-effective methodology was proposed for the prediction of the evolution of seabed level in the long-term erosion and specific points along a sandy beach were assessed. The obtained numerical results showed a fair agreement with the real status of the examined coast.

6.2 Future research directions

In this chapter, some suggestions for future research on different directions related with the scopes of the present thesis are summarized.

- Construction of trivariate statistical models for metocean data, including linear and directional variables, for the full description of sea states and wind conditions. Such models are useful for the selection of the most suitable energy device and its optimum design and arrangement (in case of arrays) at a candidate site.

- Quantitative assessment of the impact of climate change on the energy extraction of marine renewable energy sources to understand how the availability of these resources will be influenced. For this purpose, either hindcast or climate models can be used while estimates of changes for the directional variables should also be included.

- Identification of outliers in circular data through various statistics and investigation of robust estimators that are extended in the circular-circular regression model. The results of this thesis indicate that robust estimators applied in linear variables perform consistently better that the classical regression methods; hence, studying further the influence of outliers in directional data would be of considerable interest.

- As regards the directional extreme value model, it would be interesting to examine an alternative model to Fourier series expansion for expressing smoothly the periodicity of the parameters in terms of direction. Moreover, the consideration of a threshold that is directionally varying would also be meaningful while the effects of selecting various numbers of sectors, either equiangular or not, for the independent fits deserve a thorough investigation.

- Implementation of the methodology proposed for the estimation of the seabed level with simultaneous reduction of the computational time to another coastal site. Preferably, this site

should have more complex topographical features that allow a wider range of wave directions affecting the morphological conditions while the availability of *in situ* measurements of seabed is important for validation purposes. Furthermore, the proposed methodology could be enhanced. For instance, additional factors that influence sediment transport and seabed level, such as currents, could be included while the impacts of a finer resolution of the involved wave parameters during the discretization process along with a longer reference time series could be also analysed.

PART IV

APPENDICES

Appendix A Descriptive statistics for circular variables

For the representation of circular data two coordinates systems are used: i) the rectangular (or Cartesian) coordinate system, representing the location of a point Z on a plane by a pair of coordinates (x_Z, y_Z), and ii) the polar coordinate system, where the distance r between the origin and the point Z is needed along with the angle θ formed by the reference line and a line going through point; see also Figure A-1.
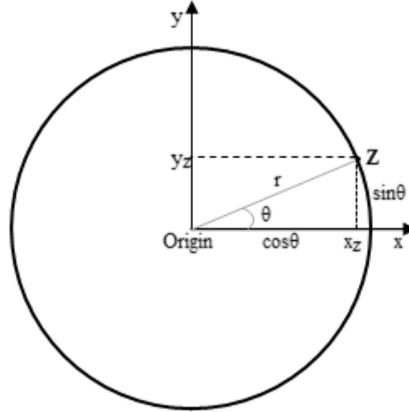
Figure A-1. Graphical representation of rectangular and orthogonal coordinate systems.

The two coordinate systems are related with each other by the following equations:

$$x_Z = r\cos\theta, y_Z = r\sin\theta. \tag{A.1}$$

Since the direction is only of interest, it is considered that we work on a unit circle (i.e. with $r = 1$ centred at the origin). Hence, Eq. *(A.1)* simply becomes

$$x_Z = \cos\theta, y_Z = \sin\theta. \tag{A.2}$$

Equivalently, another representation for circular data is achieved through complex number. The corresponding relation is $z = \exp(i\theta) = \cos\theta + i\sin\theta$, where $i = \sqrt{-1}$.

A.1   Measures of location

Let us assume a circular random variable Θ and $\theta_1, \dots, \theta_n$ a set of circular observations with n the total number of observations. The calculation of the circular mean direction is based on the trigonometric functions sine and cosine of the circular observations and is obtained by working with polar coordinates. The corresponding expression is given by:

$$m_\Theta = \bar{\theta} = \operatorname{atan2}(S/C) = \begin{cases} \operatorname{atan}(S/C), & C > 0, S \geq 0 \\ \operatorname{atan}(S/C) + 2\pi, & C \geq 0, S < 0 \\ \operatorname{atan}(S/C) + \pi, & C < 0 \\ \pi/2, & C = 0, S > 0 \\ -\pi/2, & C = 0, S < 0 \\ \text{undefined}, & C = 0, S = 0, \end{cases} \tag{A.3}$$

where $C = \sum_{j=1}^n \cos\theta_j$ and $S = \sum_{j=1}^n \sin\theta_j$. When $C = S = 0$ (and the sample size is even), the circular data are uniformly/evenly distributed or have a cyclic structure over the unit circle.

Mean direction reflects the centre of a (unimodal) set of directions and is independent of the choice of zero direction and the sense of rotation.

The quantity $R = \sqrt{C^2 + S^2} \geq 0$ is called resultant length and is associated with mean direction by the equations

$$\cos\bar{\theta} = C/R, \sin\bar{\theta} = S/R. \tag{A.4}$$

Apart from the mean direction (Eq. (A.3)), an additional measure of location is the sample median direction $\tilde{\theta}$. This measure is defined by any angle ω for which the circular data are equally divided in the arc $[\omega, \omega + \pi]$ and the majority of the data points lie closer to ω rather than the antipodal point (i.e. $\omega + \pi$). Just as in linear data, when the sample size n is odd, the median direction is one of the data points; otherwise, it is the mean value of two data points. It is obvious that the median is not uniquely defined therefore the interpretation of the results should be made with caution.

A.2   Measures of concentration, dispersion and circular distance

The mean resultant length $\bar{R} = R/n, \bar{R} \in [0, 1]$, is considered as a measure of concentration for unimodal circular data. In general, values of $\bar{R}$ close to 1 indicate that the circular data are highly concentrated about the mean direction. When $\bar{R} = 1$, it is implied that all the sample observations coincide. On the other hand, $\bar{R} = 0$ does not necessary imply that the data are evenly spread around the circle, as already mentioned above.

The circular variance, a measure of dispersion, is defined

$$V_\Theta = 1 - \bar{R}, V \in [0, 1]. \tag{A.5}$$

The circular standard deviation is given by

$$s_\Theta = \{-2\log(1 - V_\Theta)\}^{1/2}, s \in [0, \infty). \tag{A.6}$$

Another approximation for circular standard deviation, when V is small, is

$$s_\Theta \simeq \sqrt{2V_\Theta} = \sqrt{2(1 - \bar{R})}. \tag{A.7}$$

The sample circular dispersion is defined by

$$\hat{\delta} = \frac{1 - \bar{R}_2}{2\bar{R}^2}, \tag{A.8}$$

where $\bar{R}_2$ is the sample mean resultant length of $2\theta_1, \dots, 2\theta_n$.

A related definition is the circular distance measure between any two points θ, ω on the circumference in terms of arc lengths is

$$d = \min(\theta - \omega, 2\pi - (\theta - \omega)) = \pi - |\pi - |\theta - \omega||, d \in [0, \pi]. \tag{A.9}$$

The corresponding dispersion between a set angles $\theta_1, \dots, \theta_n$ and an angle ω is found by

$$d(\omega) = \frac{1}{n}\sum_{j=1}^{n}\left\{\pi - \left|\pi - \left|\theta_j - \omega\right|\right|\right\}. \tag{A.10}$$

The minimization of the above function yields the sample median direction.

An alternative expression of circular distance is given in terms of the cosine function as follows

$$d^* = 1 - \cos(\theta, \omega), d^* \in [0, 2]. \tag{A.11}$$

Skewness and kurtosis are expressed in terms of the second central sine and cosine moments, respectively, as follows

$$\hat{s} = \frac{(1/n)\sum_{j=1}^{n}\sin 2(\theta_j - \bar{\theta})}{(1-\bar{R})^{3/2}} = \frac{\bar{R}_2\sin(\bar{\theta}_2 - 2\bar{\theta})}{(1-\bar{R})^{3/2}}, \text{ and} \tag{A.12}$$

$$\hat{k} = \frac{(1/n)\sum_{j=1}^{n}\cos 2(\theta_j - \bar{\theta})}{(1-\bar{R})^{3/2}} = \frac{\bar{R}_2\cos(\bar{\theta}_2 - 2\bar{\theta}) - \bar{R}^4}{(1-\bar{R})^2}. \tag{A.13}$$

In general, positive values of kurtosis imply peaked distributions and positive values of skewness suggest that the sample data are skewed in the clockwise direction. Values of the above measures around zero suggest symmetric distributions.

A.3   Measures of correlation

The statistical association between a linear random variable X and a circular variable Θ can be quantified through the linear–circular correlation coefficient $r_{X\Theta}^2$, which is defined as follows:

$$r_{X\Theta}^2 = \frac{r_{Xc}^2 + r_{Xs}^2 - 2r_{Xs}r_{Xc}r_{cs}}{1 - r_{cs}^2}, \tag{A.14}$$

where

$$\begin{cases} r_{Xc} = \rho[(x_1, \cos\theta_1), (x_2, \cos\theta_2), .., (x_n, \cos\theta_n)] \\ r_{Xs} = \rho[(x_1, \sin\theta_1), (x_2, \sin\theta_2), .., (x_n, \sin\theta_n)] \\ r_{cs} = \rho[(\cos\theta_1, \sin\theta_1), (\cos\theta_2, \sin\theta_2), .., (\cos\theta_n, \sin\theta_n)], \end{cases} \tag{A.15}$$

with ρ denoting the Pearson product-moment correlation between X and Θ.

The statistical correlation coefficient between two directional random variables Θ, Φ are estimated

$$r_{\Theta\Phi}^2 = \frac{(r_{cc}^2 + r_{cs}^2 + r_{sc}^2 + r_{ss}^2) + 2R_{cs}r_1r_2 - 2R_{cs}r_1 - 2R_{cs}r_2}{(1-r_1^2)(1-r_2^2)}, \tag{A.16}$$

where

$$\begin{cases} R_{cs} = r_{cc}r_{ss} + r_{cs}r_{sc} \\ r_1 = \rho[(\cos\theta_1, \sin\theta_1), (\cos\theta_2, \sin\theta_2), ..., (\cos\theta_n, \sin\theta_n)] \\ r_2 = \rho[(\cos\phi_1, \sin\phi_1), (\cos\phi_2, \sin\phi_2), ..., (\cos\phi_n, \sin\phi_n)]. \end{cases} \tag{A.17}$$

A.4   Sample trigonometric moments

The p −th trigonometric moment about the zero direction is

$$m'_p = \frac{1}{n}\sum_{j=1}^n e^{ip\theta_j} = \frac{1}{n}\sum_{j=1}^n \left(\cos p\theta_j + i\,\sin p\theta_j\right) = \bar{C}_p + i\bar{S}_p, \tag{A.18}$$

where $p = 0, \pm 1, \pm 2, ...$, $\bar{C}_p = \frac{1}{n}\sum_{j=1}^n \cos p\theta_j$ and $\bar{S}_p = \frac{1}{n}\sum_{j=1}^n \sin p\theta_j$. Also note that $\bar{C}_{-p} = \bar{C}_p$ and $\bar{S}_{-p} = -\bar{S}_p$.

The polar representation of m'_p, when $\bar{R}_p > 0$, is

$$m'_p = \bar{R}_p e^{i\bar{\theta}_p} = \bar{R}_p\left(\cos\bar{\theta}_p + i\sin\bar{\theta}_p\right), \tag{A.19}$$

where $\bar{R}_p$ is the sample mean resultant length of $p\theta_1, ..., p\theta_n$ and $\bar{\theta}_p$ is the sample mean direction of $p\theta_1, ..., p\theta_n$.

From Eqs.(A.4), (A.18) and (A.19) it follows that

$$\bar{C}_p = \bar{R}_p \cos\bar{\theta}_p \text{ and } \bar{S}_p = \bar{R}_p \sin\bar{\theta}_p. \tag{A.20}$$

As regards the p −th trigonometric moment about the mean direction, the corresponding relation is

$$m_p = \bar{c}_p + i\bar{s}_p, \tag{A.21}$$

where $\bar{c}_p = \frac{1}{n}\sum_{j=1}^n \cos p\left(\theta_j - \theta\right)$ and $\bar{s}_p = \frac{1}{n}\sum_{j=1}^n \sin p\left(\theta_j - \bar{\theta}\right)$.

Appendix B Statistical analysis for metocean climate modelling

In the context of metocean climate modelling, various statistical measures can be applied for the description of the variables of interest. The most common measures are the low-order statistical moments (e.g. mean value, standard deviation) that can reveal different features of the examined phenomenon when estimated for different time scales (e.g. monthly, annual, decadal etc.) if the available time series permits such estimations; see Appendix B.1. Additional informative parameters are the measures of variability and correlation; for instance, mean annual and inter-annual variability, defined in Appendix B.2, are important measures in the climate analysis and a variety of applications such as offshore engineering industry. Moreover, the relation between linear and directional variables can be examined in terms of correlation coefficients, as described in Appendix B.3, while evaluation metrics presented in Appendix B.4, are used to quantify the quality of the regression models. In Appendix B.5, the coefficient of determination (for the univariate case) is provided, which is widely used when a number of distributions are compared in terms of suitability, and in Appendix B.6, the Mann-Kendall test, used for monotonic trend analysis, is briefly described.

B.1 Descriptive statistics for different time scales

The following notation and equations correspond to the linear variable X. The directional variable and its derived parameters follow the same nomenclature; however, in this case, the statistical parameters are calculated by using directional statistics. Some background theory of directional statistics is provided in Appendix A. Wherever necessary, the explicit notation of the directional parameters is provided.

Let denote the annual mean value of a linear variable X for a particular year j by

$$m_{u,Y=j} = \frac{1}{N} \sum_{i=1}^{N} x_i, \qquad (B.1)$$

where N is the total number of the p −hour intervals, where usually $p = 1,3,6$, for year j and x_i is the time series of the linear variable. The sequence of the above annual mean values for a series of years is denoted by $m_{x,Y}(j), j = 1,2,\dots,J$, wherefrom the mean annual value $m_{x,Y}$ is estimated as follows:

$$m_{x,Y} = \frac{1}{J} \sum_{j=1}^{J} m_{x,Y}(j). \qquad (B.2)$$

The monthly mean wind speed for a particular year j and month m, denoted by $m_{u,Y=j,M=m}$, is obtained

$$m_{u,Y=j,M=m} = \frac{1}{K} \sum_{i=1}^{K} x_i, \qquad (B.3)$$

where K is the total number of the p-hour intervals for month m of year j. The sequence of the above monthly mean values for a series of years is denoted by $m_{x,Y,M}(j,m), j = 1,2,\dots,J, m = 1,2,\dots,M$. The mean monthly value for a particular month m, $m_{x,M=m}$ can be estimated from an appropriately selected subsequence of $m_{x,Y,M}(j,m)$, as follows:

$$m_{u,M=m} = \frac{1}{J}\sum_{j=1}^{J} m_{x,M=m}(j), \text{ for } m = 1,2,\dots,12, \tag{B.4}$$

where $m_{x,M=m}(j), j = 1,2,\dots,J$, denotes the sequence of monthly mean values for the particular month m for the entire series of years.

B.2   Variability measures

The coefficient of variation is used as a relative measure of the dispersion of data points around the mean. In particular, the mean annual variability, denoted as CV (also referred to as MAV), provides a measure of variability of the linear variable within each year and is estimated by

$$\text{CV} = \frac{1}{J}\sum_{j=1}^{J} \frac{Sx_{u,Y}(j)}{m_{x,Y}(j)}, \tag{B.5}$$

where $S_{u,Y}(j)$ is the standard deviation of the linear variable for the year $j, j = 1,2,\dots,J$ (Stopa et al., 2013).

The inter-annual variability, denoted as IAV, provides an indication of the variability from year-to-year of the linear variable. It is defined as the ratio of the standard deviation $S_{m_{x,Y}(j)}$ of the annual mean value sequence $m_{x,Y=j}$, to the overall mean wind speed value m_x, i.e.

$$\text{IAV} = \frac{S_{m_{x,Y}(j)}}{m_x}. \tag{B.6}$$

B.3   Correlation measures

The statistical correlation coefficient of a series of pairs (x_i, y_i), $i = 1,2,\dots,n$ of two linear random variables X, Y (linear-linear) can be estimated by

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}. \tag{B.7}$$

The corresponding linear-directional and directional-directional correlation coefficients are defined in Appendix A.3.

B.4   Evaluation metrics for regression (linear and directional) models

In the context of evaluating the performance of different regression/calibration models, the following statistical measures can be applied as regards the linear variables:

the bias (BIAS),

168

$$\text{BIAS} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - x_i), \tag{B.8}$$

the root mean square error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{y}_i)^2}, \tag{B.9}$$

the mean absolute error (MAE)

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - x_i|, \tag{B.10}$$

and the scatter index (SI):

$$\text{SI} = \frac{\text{RMSE}}{\overline{x}}, \tag{B.11}$$

where n is the sample size of the dataset, x_i is the i−th value of the reference linear variable, $\hat{y}_i$ is the i−th value of the calibrated linear variable and $\overline{x}$ is the mean value of the reference linear variable. Clearly, the above-defined statistical measures are proportional (in various forms) to the error (difference) between the corrected and the reference linear variable. Specifically, MAE takes into consideration the sum of the absolute errors, RMSE and SI the sum of the errors squared, while BIAS the positive and negative value of this difference. Though, it should be noted that RMSE (and SI) is more sensitive to the presence of outliers than MAE, since large errors are biased towards outliers; see also Hyndman and Koehler (2006). Let us note that MAE, RMSE and SI comprise stricter and more realistic control criteria than BIAS, since the latter neutralizes the foregoing differences by definition, meaning that positive differences can be offset by negative ones. The quality of a calibration performance is characterized as 'good', if the values of the applied statistics are as close as possible to zero. Let us also remind that the values of BIAS, RMSE and MAE take the units of the variable under examination.

As regards the directional variables the following statistical measures can be applied:

the bias (BIAS),

$$\text{BIAS} = \bar{\theta} - \bar{\phi}, \tag{B.12}$$

the mean circular absolute error (MCAE) (Jing-Jing et al., 2014),

$$\text{MCAE} = \frac{1}{n}\sum_{i=1}^{n}|d(\theta_i, \phi_i)|, \tag{B.13}$$

where d is obtained from Eq. (A.9),

the root mean error (RME) (Karathanasi et al., 2016)

$$RME = \sqrt{-2\ln\left(\frac{1}{n}\sum_{i=1}^{n}\left|\cos\frac{\theta_i - \phi_i}{2}\right|\right)},$$ (B.14)

and finally, a modification of the overall estimated circular prediction bias proposed by (SenGupta et al. 2013), the mean relative bias (MRB),

$$MRB = \frac{1}{n}\sum_{i=1}^{n}\sin\left(\frac{\theta_i - \phi_i}{2}\right),$$ (B.15)

where $\bar{\theta}$ and $\bar{\phi}$ are the sample mean directions. If the values of BIAS, MCAE, RME and MRB are close to zero, then the calibration performance is good. Let us also remind that the values of BIAS and MCAE are in radians.

B.5 Goodness-of-fit testing for univariate distributions

The coefficient of determination is given by

$$R_{a,1}^2 = \frac{\sum_{i=1}^{n}\left(\hat{F}_i - \overline{F}\right)^2}{\sum_{i=1}^{n}\left(\widetilde{F}_i - \overline{F}\right)^2 + \sum_{i=1}^{n}\left(F_i - \widetilde{F}_i\right)^2},$$ (B.16)

where $\widetilde{F}$ estimate is obtained from the probability model, F_i, $i = 1, \dots, n$, (where n denotes the sample size) are obtained by using the Weibull plotting positions $F(x_i) = i/(n+1)$, $i = 1, \dots, n$, and $\overline{F} = (1/n)\sum_{i=1}^{n}\widetilde{F}_i$. The Weibull plotting positions were selected since they provide unbiased estimates of the observed cumulative probabilities regardless of the underlying distribution.

B.6 Mann-Kendall test

The Mann-Kendall test is a non-parametric test, frequently used to detect the existence of monotonic (upward or downward) trends in time series; see, e.g. Hipel and McLeod (1994). The test is based on the correlation between ranks of a time series and their order, instead of the actual values of the series, and it is less sensitive to the presence of outliers. The null hypothesis H0 is that the data come from a population of independent and identically distributed variables. An important advantage of Mann-Kendall test is that it is distribution-free in contrast, for example, to the regression slope test, where the residuals are assumed to be normally distributed. On the other hand, the examined data should not be serially correlated in order for the estimated p values to be correct. The Mann-Kendall test statistic is calculated as follows:

$$S = \sum_{k=1}^{n-1}\sum_{j=k+1}^{n} sgn(X_j - X_k),$$ (B.17)

where

$$sgn(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{if } x = 0 \\ -1, & \text{if } x < 0. \end{cases}$$ (B.18)

The variance of S (in the general case where ties are present) is given as follows:

$$\sigma_S^2 = \frac{1}{18}\left[n(n-1)(2n+5) - \sum_{q=1}^{Q} t_q(t_q-1)(2t_q+5)\right], \tag{B.19}$$

where n denotes the time series length, Q is the number of tied groups and t_q is the number of observations in the $q-$th tied group. The final test statistic Z results from the following transformation of S:

$$Z = \begin{cases} (S-1)/\sigma_S, & \text{if } S > 0 \\ 0, & \text{if } S = 0 \\ (S-1)/\sigma_S, & \text{if } S < 0. \end{cases} \tag{B.20}$$

Positive (negative) values of Z suggest an upward (downward) trend. As the sample size becomes larger, the test statistic Z follows approximately the Gaussian distribution.

B.7 Mardia-Wheeler-Watson test

Mardia-Wheeler-Watson test is a non-parametric procedure that can be applied to samples of circular data in order to test the hypothesis that the examined samples have identical circular distributions regarding mean direction, circular variance or both.

Let θ_1 and θ_2 be two independent random samples with sample size n_1 and n_2, from populations with continuous circular distributions $P_1(\theta)$ and $P_2(\theta)$, respectively. In order to test the null hypothesis

$$H_0: P_1(\theta) = P_2(\theta), \tag{B.21}$$

the circular ranks (uniform scores) of the combined sample are calculated; let $(r_1, \dots, r_{n1})$ be the ranks of the directions of the first sample. Mardia-Wheeler-Watson test statistic is based on the criterion

$$W = \frac{2(N-1)(C_j^2 - S_j^2)}{n_1 n_2}, \tag{B.22}$$

where $C_j = \sum_{i=1}^{n_j} \cos(2\pi r_i/N)$, $S_j = \sum_{i=1}^{n_j} \cos(2\pi r_i/N)$ with j referring to sample either 1 or 2, and $N = n_1 + n_2$. In order to apply efficiently this approximation, $N > 17$ as proposed by Batschelet (1981) or $N \geq 10$ as proposed by Fisher (1993), which is also similar to the proposition of Mardia and Spurr (1973). Furthermore, circular data of samples should not be tied (i.e. equal numerical values in the combined sample) or the two sample dispersions should not be very different (Batschelet, 1981). It has been shown that W approaches a χ^2-distribution with two degrees of freedom for large N. The null hypothesis is rejected for large values of W ($>\chi_{\alpha,2}^2$) (Wheeler and Watson, 1964; Mardia, 1967; Batschelet, 1981).

Appendix C Extremes

C.1   Parameter estimation for GP distribution

The GP parameters are usually estimated by the maximum likelihood method (del Castillo and Serra, 2015; Grimshaw, 1993), the probability weighted moments introduced by Greenwood et al. (1979), the method of moments (Hosking and Wallis, 1987) and the elemental percentile method based on a two-stage procedure proposed by Castillo and Hadi (1997).; see also the extensive study of Bermudez and Kotz (2010) on this issue.

The most popular method among the estimators is the maximum likelihood (ML) method. The likelihood function is the joint pdf of a random sample $x_1, \dots x_n$ from a distribution with pdf $f(x_i; \theta)$ as a function of θ, and is defined as

$$L(\theta; \mathbf{x}) \equiv L(\theta) = \prod_{i=1}^{n} f(x_i; \theta), \theta \in \Omega, \tag{C.1}$$

where θ is the unknown parameter (in a set Ω) on which f depends, and $\mathbf{x} = (x_1, \dots, x_n)^{\mathrm{T}}$. Since the natural logarithm is a monotonically increasing function, for convenience in the calculations, the natural logarithm of the likelihood function, $\ell(\theta) = \ln L(\theta) = \sum_{i}^{n=1} f(x_i; \theta)$, called log-likelihood function, is used to estimate the values of the parameters that maximize this function, denoted by $\hat{\theta}$. Subsequently, $\hat{\theta}$ is the maximum likelihood estimator of θ, which can be obtained by differentiating the log-likelihood function with respect to θ and solving the following equation:

$$\frac{\partial \ln L(\theta)}{\partial \theta} = 0. \tag{C.2}$$

The log-likelihood function of the GP distribution is given by

$$\ell(\sigma_u, \xi) = \begin{cases} -n_u \log \sigma_u - \left(1 + \dfrac{1}{\xi}\right) \displaystyle\sum_{j=1}^{n_u} \log \left(1 + \dfrac{\xi y_j}{\sigma_u}\right), \xi \neq 0, 1 + \dfrac{\xi y_j}{\sigma_u} > 0 \\[2em] -n_u \ln \sigma_u - \dfrac{1}{\sigma_u} \displaystyle\sum_{j=1}^{n_u} y_j, \qquad\qquad\qquad \xi = 0. \end{cases} \tag{C.3}$$

Assuming that $\ell(\cdot)$ is differentiable, the ML estimator $\hat{\lambda} = (\hat{\sigma}_u, \hat{\xi})$ for the unknown parameters $\lambda = (\sigma_u, \xi)$ is obtained by maximizing Eq. (C.3). The maximization problem is solved numerically using optimization/iterative methods, such as Newton-Raphson method and Expectation-Maximization algorithm, since no explicit solution exists for the equations derived after differentiating the above equation.

Note that ML estimators do not exist for $\xi > 1$ since log-likelihood becomes infinite and have higher efficiency when ξ is close to zero. In general, for $\xi \leq 1$, ML estimators exist.

Appendix D Numerical model: MIKE21 by DHI

The numerical modelling package that was used for the purposes of this thesis is MIKE 21/3 Coupled Model Flexible Mesh (hereafter MIKE21 CFM) developed by DHI Group (DHI, 2016a). The two-dimensional depth-averaged flow (2DH) simulations are based on a cell centred finite volume method with an unstructured mesh for the more accurate representation of complex coastal areas while the basic principles that are applied are the conservation of fluid momentum, wave energy and mass (fluid and sediment). This modelling suite can be implemented for various hydraulic phenomena in lakes, rivers, estuaries, bays, coastal areas and seas through a dynamic modelling system; in the context of this thesis, it combines three different modules: (i) the hydrodynamic (HD) module; (ii) the spectral wave (SW) module, and; (iii) the sand transport (ST) module.

The general structure behind this model is first to analyse the available data as concerns initial bathymetry, flow patterns, sediment composition etc. Then, the core computational components of MIKE21 CFM are the first two modules, which simulate the mutual interaction between currents and waves using a dynamic coupling for the determination of the hydraulic conditions for the initial and subsequent situations. The results of these models are used as input for the additional dynamic coupling that includes the third module, which gives a full feedback of the seabed level changes on waves and flow calculations resulting in a new bathymetry until the predefined final time period is reached. Let us note that the sediment transport rates and morphological changes are calculated simultaneously with the hydrodynamics. The simulation of flows and transports in marine, coastal and estuarine areas is based on a flexible mesh approach. The quality of the available input data for all the above modules (flow velocity, wave parameters, grain composition) is of critical importance as well as the data that will be used for calibrating the model results.

D.1 Hydrodynamic (HD) module

The hydrodynamic model solves the 2D incompressible Reynolds averaged Navier-Stokes equations under the Boussinesq simplifying approximation and the hypothesis of hydrostatic pressure. The continuity equation (in horizontal Cartesian coordinates) over water depth $h = \eta + d$, with η denoting the surface elevation and d the still water depth, is the following:

$$\frac{\partial h}{\partial t} + \frac{\partial h\bar{u}}{\partial x} + \frac{\partial h\bar{v}}{\partial y} = hS_{\text{HD}}, \tag{D.1}$$

where t is time, $h\bar{u}$ and $h\bar{v}$ denote the depth-averaged values of the velocity components in the $x-$ and $y-$direction, respectively, and S_{HD} is the magnitude of discharge due to point sources. The momentum equations for the $x-$ and $y-$component are the following:

$$\frac{\partial h\bar{u}}{\partial t} + \frac{\partial h\bar{u}^2}{\partial x} + \frac{\partial h\bar{v}\bar{u}}{\partial y} = f\bar{v}h - gh\frac{\partial \eta}{\partial x} - \frac{h}{\rho_0}\frac{\partial p_a}{\partial x} - \frac{gh^2}{2\rho_0}\frac{\partial \rho}{\partial x} + \frac{\tau_{sx}}{\rho_0} - \frac{\tau_{bx}}{\rho_0}$$
$$- \frac{1}{\rho_0}\left(\frac{\partial s_{xx}}{\partial x} + \frac{\partial s_{xy}}{\partial y}\right) + \frac{\partial(hT_{xx})}{\partial x} + \frac{\partial(hT_{xy})}{\partial y} + hu_s S_{\text{HD}}, \tag{D.2}$$

and

$$\frac{\partial h\bar{v}}{\partial t} + \frac{\partial h\bar{u}\bar{v}}{\partial x} + \frac{\partial h\bar{v}^2}{\partial y} = -f\bar{u}h - gh\frac{\partial \eta}{\partial y} - \frac{h}{\rho_0}\frac{\partial p_a}{\partial y} - \frac{gh^2}{2\rho_0}\frac{\partial \rho}{\partial y} + \frac{\tau_{sy}}{\rho_0} - \frac{\tau_{by}}{\rho_0}$$
$$- \frac{1}{\rho_0}\left(\frac{\partial s_{yx}}{\partial x} + \frac{\partial s_{yy}}{\partial y}\right) + \frac{\partial(hT_{xy})}{\partial x} + \frac{\partial(hT_{yy})}{\partial y} + hv_s S_{\text{HD}}, \tag{D.3}$$

where f is the Coriolis parameter, g is the gravitational acceleration, p_a is the atmospheric pressure, ρ_0 is the reference density of water, ρ is the water density, s_{ij} is the radiation stresses, τ_s is the surface wind stress and τ_b is the bottom stress, and T_{ij} is the lateral stresses and u_s, v_s are the velocity components at which the water is discharged.

A cell-centred finite volume technique is used for the spatial discretization of the domain. An approximate Riemann solver, known as Roe's scheme (Roe, 1981), is used for the computation of the convective fluxes. The solution of Eqs. (D.2) and (D.3) result in the values of water particle velocities and current components that are responsible for the sediment transport in the coastal zone. Smagorinsky formulation is used to represent horizontal eddy viscosity while bottom friction can be specified by the Manning's roughness coefficient.

D.2 Spectral wave (SW) module

The spectral wave module is a third-generation spectral wind-wave generation model that simulates the growth, decay and transformation of wind-generated waves and swells both in offshore and coastal regions, and is based on unstructured meshes (in the geographical domain). The above simulations are based on the conservation equation of the wave action expressed by Eq. (5.1). In the source term S, the following physical phenomena are included:

- wave generation and growth by wind action S_{in}, proposed by the quasi-linear theory developed by Janssen in a series of studies (Janssen, 1989; Janssen, 1991; Janssen et al., 1989) as regards wind and wave interaction;
- wave energy transfer due to non-linear wave-wave interaction S_{nl}, using the Discrete Interaction Approximation (DIA) of Hasselmann et al. (1985);
- dissipation of wave energy due to white-capping S_{wc}, proposed by Hasselmann (1974) and tuned according to Janssen (1989) and Janssen (1992);
- dissipation of wave energy due to bottom friction S_{bf}, based on the approach of Johnson and Kofoed-Hansen (2000), which takes into consideration wave and sediment properties;
- dissipation of wave energy due to wave breaking S_{wb}, based on the breaking model of Battjes and Janssen (1978), and Eldeberky and Battjes (1996).

Wind forcing and diffraction can also be included in the wave model. For the discretization of the governing equations in the geographical and spectral space, a cell-centred finite volume formulation is used by subdividing the continuum into non-overlapping elements, while a multi-sequence explicit scheme is applied for the wave propagation, and a fractional step method is implemented for the time integration, where an explicit method is used for solving the propagation step. Two different formulations are included in this module: i) the fully spectral formulation, suitable for near-shore applications, and ii) the directional decoupled parametric formulation, mostly used for offshore wave modelling.

D.3 Sediment transport (ST) module

The computed flow and wave fields are used as input for the sediment transport model. The modelling of non-cohesive sediment (i.e. sand) transport fields for the calculation of seabed

level changes and sediment transport rates can be obtained by using the theory of combined waves and currents, including the wave breaking effect. This module can be applied in coastal regions (at a local or regional scale), such as estuaries, tidal inlets and coasts, as well as in coastal structures, such as harbours.

The total sediment transport q_{tot} is defined by the bed load transport q_b (i.e. load that is in continuous contact with the seabed during transport) and the sediment transport in suspension q_s (i.e. load that is moving without being in contact with the seabed due to the turbulent flow). Wash load (i.e. finer suspended material than that of the seabed that remain in permanent suspension) is considered negligible in the coastal environment and is not included in this module.

As regards the bed load transport, it is calculated from the instantaneous Shields parameter according to the model that was proposed by Engelund and Fredsøe (1976). Based on the equilibrium of agitating and stabilizing forces on a sediment particle, Engelund and Fredsøe (1976) proposed the following non-dimensional form that proved to be accurate for fine to medium sediments:

$$\Phi_B = 5p\left(\sqrt{\theta} - 0.7\sqrt{\theta_c}\right), \tag{D.4}$$

where p represents the probability that a certain fraction of the sediment particles are in movement in a single layer and θ_c is the critical Shields parameter equal to 0.045. The probability p is defined by:

$$p = \left(1 + \left(\frac{\frac{\pi}{6}\mu_d}{\theta - \theta_c}\right)^4\right)^{-1/4}, \tag{D.5}$$

where μ_d is a dynamic friction coefficient, estimated at 0.51.

In contrast with bed load that responds instantaneously with the flow, the suspended sediment transport is characterized by a phase-lag as regards its transport and it is the result of the product of the time-averaged instantaneous flow velocities u and the instantaneous sediment concentration c by integrating over the local water depth h:

$$q_s = \frac{1}{T}\int_0^T \int_{2d}^h (uc)\,dz\,dt, \tag{D.6}$$

where d is the characteristic grain dimeter, usually equal to the median grain diameter d_{50}.

The sediment transport rates are found by linear interpolation from a sediment transport table, which is calculated prior to the main model run in order to speed up the calculations, using an intra wave force description. The integrated momentum approach of Fredsøe (1984) is used for the time integration of the boundary layer. Based on equilibrium sediment transport method, the values of this table are derived from a quasi-3D numerical model (STPQ3D), which calculates in the two horizontal dimensions (longshore and cross-shore) time-averaged and instantaneous hydrodynamic flow conditions that drive sediment transport algorithms with an one-dimensional flow velocity profile model as regards the vertical direction; see Elfrink et al. (1996). The calculation of the sediment transport rates is based on an intra wave force balance description that takes into account shear stresses on bottom and wave breaking among others.

D.4   Morphology

The determination of the seabed level change at each element is based on the Exner equation (sediment continuity equation) that is written (in horizontal Cartesian coordinates) as follows:

$$-(1-n)\frac{\partial z_b}{\partial t} = \frac{\partial S_{x,\text{ST}}}{\partial x} + \frac{\partial S_{y,\text{ST}}}{\partial y} - \Delta S_{\text{ST}}, \tag{D.7}$$

where n is the bed porosity, z_b is the seabed level, t is time, $S_{x,\text{ST}}$, $S_{y,\text{ST}}$ are the total load transport in the x, y −direction, respectively, and ΔS_{ST} is the sediment source/sink rate. For an equilibrium description, the source/sink term is set to zero, unless lateral sediment supply is considered.

Based on the seabed level change rates, the seabed level is updated for every N[th] HD-time step, where N is a time step factor defined by the user. The new values of the seabed level are obtained by solving the above differential equation with a forward-in-time difference scheme as follows:

$$z_{new} = z_{old} + \frac{1}{1-n}\frac{\partial z}{\partial t}\Delta t_{\text{HD}}. \tag{D.8}$$

The morphodynamics are fully integrated with above-mentioned modules allowing the seabed level changes to provide input for the flow and wave fields so that they are adapted to the new bathymetrical state.

Appendix E Datasets

Along the thesis, three data sources have been considered for modelling the wind and wave characteristics: i) *in situ* measurements using oceanographic buoys; ii) satellite data, and; iii) results from numerical models. In this section, these data sources are summarised and the corresponding data sets from each type are presented. In Table E-6-1, the overall quality and features of each data source, characterized by its own strengths and shortcomings, is presented.

The highest quality and most accurate metocean data are obtained by *in situ* measurements. For example, offshore wind data can be obtained from meteorological masts, oceanographic buoys and fixed platforms and can be real-time or past measurements. In case of a marine renewable project development, it is essential to consider the acquisition of such data, which are accompanied by rather high costs due to the installation, operation, and maintenance, with the aim of providing an accurate analysis and forecasts, and mitigating involved risks. Nevertheless, *in situ* measurements suffer from data incorrectly recorded because of malfunction of the measuring device (wind sensor), defects in the power supply, errors in the data entry or during the measurement analysis process, etc. Moreover, such measurements are affected by external conditions, since these measuring devices operate in a dynamically changing environment; for example, there can be deviations in the rotation movements (i.e. roll, pitch and yaw) of the buoy due to the presence of sea waves or currents of high intensity. Despite the abovementioned deficiencies of in situ measurements, which can reinforce the presence of outliers, they have historically been considered as the primary reference data source for the validation and calibration of gridded metocean data as a result of the increased measurement accuracy; see, for example, Gower (1996).

As *in situ* measurements are scarce in space and the ocean conditions are rough, remote sensing techniques play an important role. Remotely-sensed data mainly refers to satellite observations that can be a considerable supplement to in situ measurements. The most satellite systems that are used for the quantification of metocean components through the transmission and reception of specific electromagnetic signals are insensitive to the meteorological conditions, solar illumination and day/night cycle. Nevertheless, the periodic coverage and spatial resolution remain a limiting issue for satellite products. A major step in the efficient utilization of satellite data refers to the enrichment of their corresponding spatiotemporal coverage; this can be achieved by appropriately blending different satellite (and sensor) products and generating gridded data sets by applying interpolation techniques. Although the quality of satellite data is often considered unsuitable for coastal areas due to the large biases near land-sea boundaries, (see, for example, Carvalho et al. (2014a)) yet, they can provide the basic starting point for the preliminary (comparative) offshore wind power assessment (Lizuma et al., 2013).

Numerical modelling data sets come from the numerical solution of the equations that govern the physical processes of metocean parameters in hindcast or forecast mode. In order to run such models an assimilation procedure, known as analysis, is implemented so that numerous point measurements around the world are integrated as well. Reanalysis schemes produce long-term time series that are suitable for climatological analyses. These data are available from specific providers and they cover over 30-year periods of wind and wave parameters all over the oceans with a sufficient resolution. The main limitation of numerical models is the uncertainties involved in the initial and boundary conditions (e.g. from data of low accuracy) that end up multiplying within the assimilation procedure.

Table E-6-1. Advantages and disadvantages of metocean data sources used in this thesis.

Data source	Advantages	Disadvantages
In situ measurements	• High precision measurement of physical quantities • Data provided as time series	• Gaps in recordings due to instrument malfunction • Measurements errors due to external conditions • Point measurement coverage • Usually of short duration • Expensive and not easily accessible
Satellite data	• Global spatial coverage (of coarse resolution) • Usually of long duration • The next more reliable source after in situ measurements	• Questionable quality at the land-sea boundary • Errors when converting the satellite's original measurements to the corresponding values of physical quantity and in interference patterns • Periodicity of satellite tracks
Data from numerical models	• Global spatial coverage • Long duration • Easily accessible and usually free of charge • Data provided as time series	• Uncertain results due to the model errors (e.g. parameterization errors, low accuracy of initial and boundary conditions) • Subject to errors when complex topography is represented with a coarse spatial resolution

E.1 *In situ* measurements

In the Greek Seas, a network of eleven oceanographic buoys, deployed in deep water locations, operates within the framework of the POSEIDON marine monitoring and forecasting system since 2000 under HCMR (Soukissian and Chronis, 2000). Each buoy is equipped with meteorological and oceanographic sensors for measuring, among others, temperature, atmospheric pressure, salinity. The wind measurements are performed at 3 m height above sea surface with recording period 600 s and frequency 1 Hz, and the measurements are performed every 3 h. In this thesis, the buoy wind and wave data consist of long-term time series of wind speed, wind direction, significant wave height, wave period and wave direction for various locations in the Aegean Sea and with varying recording periods. Today, four buoys, measuring wind and wave parameters, are operating in the Aegean Sea.

In the Spanish Seas, twelve oceanographic buoys located in deep water depths provide data from measured parameters similar to the Greek buoys. The monitoring system operates under the responsibility of the Spanish Port Authority (Puertos del Estado). The measurements are made at 3 m height above sea surface with a recording interval of 1 h. The corresponding wind speed time series cover time periods varying between 5 and 18 years. In this thesis, only wind data have been analysed from the Spanish network from three buoys located in the Mediterranean part of the Spanish waters.

Wind data from both Greek and Spanish buoys were utilized in offshore wind energy applications presented in Sections 4.3 and 4.4. Furthermore, wind and wave measurements from

a Greek buoy location were used for the description of wind and wave climatology, respectively, and to validate the wave model in Section 5.4 along with an AWAC profiler.

E.2   Satellite data

The satellite data used in this thesis refer to wind speed and direction time series obtained from the Blended Sea Winds (BSW), developed by the National Climatic Data Center (NCDC) of the NOAA agency. This dataset consists of blending observations of ocean surface vector winds and wind stresses from long-term multiple satellites (up to six satellites since June 2002). In this way, a larger spatial and temporal coverage of the measurements is feasible compared to the individual satellite data sets. Specifically, NOAA has developed blended satellite products on a global basis, with 6-h temporal resolution at a spatial resolution of 0.25°×0.25° (~28 km×28 km). In 1987, there was only one satellite, while in 2000, there were more than five (SSMIF13, SSMI F14, SSMI F15, TMI, QuikSCAT, AMSR-E). The common characteristic of the blended product is that it is based on the same retrieval algorithms for all instruments involved. Regarding the sampling time intervals, they generally decrease as the considered latitudes increase. Since 2000, the sampling time intervals have decreased to less than 5 h in the tropics and higher latitudes. From the blended satellite product, the 20-yeartime series extending from January 1, 1995 to December 31, 2014 with a 6-h time window, at 10 m height above sea level was extracted and analysed. The wind directions of the BSW product originate from the NCEP/DOE (Department of Energy) Reanalysis II and are interpolated onto the BSW grid. A detailed description of BSW datasets can be found in Zhang et al. (2006). This dataset was considered in the application of Section 4.3 and can be downloaded from ftp://eclipse.ncdc.noaa.gov/pub/seawinds/ SI/uv/.

Evidently, the procedure of blending the ocean surface winds from multiple satellites into a single product results in inhomogeneous data coverage in the spatial domain; see Fig. 1 of Soukissian et al. (2017), where the temporal percentage of BSW data coverage for the Mediterranean Basin is depicted for 1995–2014. The majority of the 0.25°∘boxes are sampled more than 75% of the examined period within each 6-h time window. Specifically, the temporal percentage is satisfactory in the open sea and offshore areas of all the main sub-basins (i.e. western Mediterranean, Ionian, and Levantine Seas), where the number of measurements is more than adequate. In the Adriatic Sea and a large part of the Aegean Sea, this number is decreased, but not drastically, still permitting statistically robust assessments. Areas where the satellite temporal coverage percentage is very poor (10% of the maximum expected number of observations), and thus not acceptable for further statistical analysis, are identified in the northern and central Aegean Sea. The results referring to these areas are disregarded from the analysis. On the other hand, the data availability near the coasts may also be rather low. However, the evaluation procedure with buoy measurements showed that the collocated data sample sizes were adequate in this context.

E.3   Data from numerical models

A global atmospheric reanalysis product was used in this thesis, the ERA-Interim dataset, released by the European Centre for Medium-Range Weather Forecasts (ECMWF), with spatial resolution of approximately 80 km (or 0.75°), covering the geographical area of the Mediterranean Sea and extending from 1979 to 2014 (Dee et al., 2011). ERA-Interim is the successor of ERA-40 and stopped its update onwards in time in August 2019. It uses the ECMWF Integrated Forecasting System (IFS) Cy31r2 model and the 4-D variational data assimilation and has many enhancements compared to ERA-40 such as new wave height data from altimeters, bias correction of satellite radiance data and improved model physics. Moreover, the data quality provided by ERA-Interim is more homogeneous than its predecessor

and has additional simulate data. The quality of the ERA-Interim reanalysis wind data has been extensively verified in other works; see, for example, Alvarez et al. (2014); Carvalho et al. (2014b)).

The wave-model component of ERA-Interim is based on the WAM model, the spatial resolution is 110 km×110km, provided every 6 h, with the directional wave spectra $S(\omega, \theta)$ being discretized in 24 directions θ and 30 frequencies ω. The quality of the ERA-Interim wave data has been assessed, among others, by Stopa and Cheung (2014), and was characterized as a reliable dataset, especially for climate studies.

In Section 4.2, the offshore wind climate is analysed with this dataset. An inherent difficulty in this analysis is related to the rather low spatial resolution, which renders the obtained results less accurate near the coasts and in narrow straits and basins. However, this difficulty is not an intractable hindrance since the aim of this study is to provide a general overview of the long-term wind climatology over the Mediterranean Sea; an in-depth wind climate analysis in coastal areas requires different data in order to be successfully accomplished. Sections 4.3 and 4.4 utilized the ERA-Interim wind speed and wind direction data, respectively, while Section 4.5 included a study of the directional extreme value model based on the ERA-Interim wave data. As regards wind data, the examined time series referred to wind direction at 10 m above sea level and 6-hour time intervals while concerning wave data, the variables analysed were significant wave height and mean wave direction at the same temporal resolution. Finally, the time series of wave statistical parameters used in Section 5.5 as input for the MIKE 21 coupled model was obtained from the Mediterranean Sea Waves forecast system, which is based on the third-generation wave model WAM Cycle 4.5.4 (Günther and Behrens, 2012). The current velocity time series were obtained from the Med-currents system, whose equations are solved by an Ocean General Circulation Model based on the NEMO model (version 3.6); for more details, see Clementi et al. (2017). Both datasets can be accessed at http://marine.copernicus.eu/services-portfolio/access-to-products/.

References

Aarnes, O. J., Abdalla, S., Bidlot, J. R. & Breivik, O. 2015. Marine wind and wave height trends at different ERA-Interim forecast ranges. *Journal of Climate,* **28**(2), 819-837.

Acero, F. J., Garcia, J. A. & Gallego, M. C. 2011. Peaks-over-threshold study of trends in extreme rainfall over the Iberian Peninsula. *Journal of Climate,* **24**(4), 1089-1105.

Adams, P. N., Inman, D. L. & Lovering, J. L. 2011. Effects of climate change and wave direction on longshore sediment transport patterns in Southern California. *Climatic Change,* **109**, 211-228.

Aggarwal, C. C. 2016. *Outlier Analysis*, Springer International Publishing.

Akdağ, S. A. & Guler, Ö. 2018. Alternative Moment Method for wind energy potential and turbine energy output estimation. *Renewable Energy,* **120**, 69-77.

Akpinar, S. & Kavak Akpinar, E. 2009. Estimation of wind energy potential using finite mixture distribution models. *Energy Conversion and Management,* **50**(4), 877-884.

Alavi, O., Mohammadi, K. & Mostafaeipour, A. 2016a. Evaluating the suitability of wind speed probability distribution models: A case of study of east and southeast parts of Iran. *Energy Conversion and Management,* **119**, 101-108.

Alavi, O., Sedaghat, A. & Mostafaeipour, A. 2016b. Sensitivity analysis of different wind speed distribution models with actual and truncated wind data: A case study for Kerman, Iran. *Energy Conversion and Management,* **120**, 51-61.

Alvarez, I., Gomez-Gesteira, M., Decastro, M. & Carvalho, D. 2014. Comparison of different wind products and buoy wind data with seasonality and interannual climate variability in the southern Bay of Biscay (2000-2009). *Deep-Sea Research Part II-Topical Studies in Oceanography,* **106**, 38-48.

Amoudry, L. O. & Souza, A. J. 2011. Deterministic coastal morphological and sediment transport modeling: a review and discussion. *Reviews of Geophysics,* **49**.

Anagnostou, C., Belibassakis, K. & Karathanasi, F. 2017. Coastal erosion in the Sitia Crete Bay - Rehabilitation of the coast based on nourishment techniques as an alternative to hard work interventions *7th National Conference on Management and Improvement of Coastal Zones.* 20-22 November, Athens.

Andrews, D. G. & Mcintyre, M. E. 1978. On wave-action and its relatives. *Journal of Fluid Mechanics,* **89**(4), 647-664.

Antonutti, R., Peyrard, C., Johanning, L., Incecik, A. & Ingram, D. 2016. The effects of wind-induced inclination on the dynamics of semi-submersible floating wind turbines in the time domain. *Renewable Energy,* **88**, 83-94.

Arns, A., Wahl, T., Haigh, I. D., Jensen, J. & Pattiaratchi, C. 2013. Estimating extreme water level probabilities: A comparison of the direct methods and recommendations for best practise. *Coastal Engineering,* **81**, 51-66.

Arslan, T., Bulut, Y. M. & Yavuz, A. A. 2014. Comparative study of numerical methods for determining Weibull parameters for wind energy potential. *Renewable & Sustainable Energy Reviews,* **40**, 820-825.

Atan, R., Goggins, J. & Nash, S. 2016. A detailed assessment of the wave energy resource at the Atlantic marine energy test site. *Energies,* **9**(11).

Athanassoulis, G. A. & Belibassakis, K. A. 2002. Probabilistic description of metocean parameters by means of kernel density models 1. Theoretical background and first results. *Applied Ocean Research,* **24**(1), 1-20.

Athanassoulis, G. A., Skarsoulis, E. K. & Belibassakis, K. A. 1994. Bivariate distributions with given maringals with an application to wave climate description. *Applied Ocean Research,* **16**(1), 1-17.

Ayat, B. 2013. Wave power atlas of Eastern Mediterranean and Aegean Seas. *Energy,* **54**, 251-262.

Bücher, A. & Zhou, C. 2018. A horse racing between the block maxima method and the peak-over-threshold approach. URL: https://arxiv.org/pdf/1807.00282.pdf.

Bachynski, E. E., Kvittem, M. I., Luan, C. Y. & Moan, T. 2014. Wind-wave misalignment effects on floating wind turbines: motions and tower load effects. *Journal of Offshore Mechanics and Arctic Engineering-Transactions of the Asme,* **136**(4).

Bai, G. H., Fleck, B. & Zuo, M. J. 2016. A stochastic power curve for wind turbines with reduced variability using conditional copula. *Wind Energy,* **19**(8), 1519-1534.

Bai, Z. D., Rao, C. R. & Zhao, L. C. 1988. Kernel estimators of density function of directional data. *Journal of Multivariate Analysis,* **27**(1), 24-39.

Balakrishnan, N. & Kateri, M. 2008. On the maximum likelihood estimation of parameters of Weibull distribution based on complete and censored data. *Statistics & Probability Letters,* **78**(17), 2971-2975.

Balakrishnan, N. & Wang, J. 2000. Simple efficient estimation for the three-parameter gamma distribution. *Journal of Statistical Planning and Inference,* **85**(1-2), 115-126.

Balkema, A. A. & de Haan, L. 1974. Residual Life Time at Great Age. *Annals of Probability,* **2**(5), 792-804.

Banerjee, A., Dhillon, I. S., Ghosh, J. & Sra, S. 2005. Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research,* **6**, 1345-1382.

Barj, L., Stewart, S., Stewart, G., Lackner, M., Jonkman, J. & Robertson, A. 2014. *Wind/wave misalignment in the loads analysis of a floating offshore wind turbine: Preprint*, National Renewable Energy Lab (NREL), Golden, CO (United States).

Barnett, V. & Lewis, T. 1994. *Outliers in Statistical Data*, Wiley.

Barrett, B. E. & Gray, J. B. 1997. Leverage, residual, and interaction diagnostics for subsets of cases in least squares regression. *Computational Statistics & Data Analysis,* **26**(1), 39-52.

Bartolucci, A. A., Singh, K. P., Bartolucci, A. D. & Bae, S. 1999. Applying medical survival data to estimate the three-parameter Weibull distribution by the method of probability-weighted moments. *Mathematics and Computers in Simulation,* **48**(4-6), 385-392.

Basak, P., Basak, I. & Balakrishnan, N. 2009. Estimation for the three-parameter lognormal distribution based on progressively censored data. *Computational Statistics & Data Analysis,* **53**(10), 3580-3592.

Basile, S., Burlon, R. & Morales, F. Joint probability distributions for wind speed and direction. A case study in Sicily. In: 4th International Conference on Renewable Energy Research and Applications, 22-25 November 2015 Palermo, Italy. pp. 1591-1596.

Batschelet, E. 1981. *Circular Statistics in Biology,* New York, USA, Academic Press.

Battjes, J. A. & Janssen, J. P. F. M. Energy loss and set-up due to breaking random waves. 16th Conference on Coastal Engineering, ASCE, 1978 Hamburg, Germany. pp. 569-587.

Begueria, S. 2005. Uncertainties in partial duration series modelling of extremes related to the choice of the threshold value. *Journal of Hydrology,* **303**(1-4), 215-230.

Belibassakis, K. A. & Karathanasi, F. E. 2017. Modelling nearshore hydrodynamics and circulation under the impact of high waves at the coast of Varkiza in Saronic-Athens Gulf. *Oceanologia,* **59**(3), 350-364.

Bergillos, R. J., Rodriguez-Delgado, C. & Ortega-Sanchez, M. 2017. Advances in management tools for modeling artificial nourishments in mixed beaches. *Journal of Marine Systems,* **172**, 1-13.

Berkson, J. 1969. Estimation of a linear function for a calibration line - consideration of a recent proposal. *Technometrics,* **11**(4), 649-660.

Bermudez, P. D. & Kotz, S. 2010. Parameter estimation of the generalized Pareto distribution-Part II. *Journal of Statistical Planning and Inference,* **140**(6), 1374-1397.

Berry, W. D. 1993. *Understanding Regression Assumptions*, SAGE Publications.

Besio, G., Mentaschi, L. & Mazzino, A. 2016. Wave energy resource assessment in the Mediterranean Sea on the basis of a 35-year hindcast. *Energy,* **94**, 50-63.

Booij, N., Ris, R. C. & Holthuijsen, L. H. 1999. A third-generation wave model for coastal regions - 1. Model description and validation. *Journal of Geophysical Research-Oceans,* **104**(C4), 7649-7666.

Botev, Z. I., Grotowski, J. F. & Kroese, D. P. 2010. Kernel density estimation via diffusion. *Annals of Statistics,* **38**(5), 2916-2957.

Bowman, K. O. & Shenton, L. R. 2002. Problems with maximum likelihood estimation and the 3 parameter gamma distribution. *Journal of Statistical Computation and Simulation,* **72**(5), 391-401.

Bozzi, S., Besio, G. & Passoni, G. 2018. Wave power technologies for the Mediterranean offshore: Scaling and performance analysis. *Coastal Engineering,* **136**, 130-146.

Bretherton, F. P. & Garrett, C. J. R. 1968. Wavetrains in inhomogeneous moving media. *Proceedings of the Royal Society, England, Series A, ,* **302**(1471), 529-554.

Brown, B. M. & Chen, S. X. 1999. Beta-Bernstein smoothing for regression curves with compact support. *Scandinavian Journal of Statistics,* **26**(1), 47-59.

Burges, S. J., Lettenmaier, D. P. & Bates, C. L. 1975. Properties of the three-parameter log normal probability distribution. *Water Resources Research,* **11**(2), 229-235.

Burrows, R. & Salih, B. A. 1986. Statistical modelling of long-term wave climates. *Coastal Engineering Proceedings,* **20**, 45-56.

Callaghan, D. P., Roshanka, R. & Andrew, S. 2009. Quantifying the storm erosion hazard for coastal planning. *Coastal Engineering,* **56**(1), 90-93.

Calvo, E., Simo, R., Coma, R., Ribes, M., Pascual, J., Sabates, A., et al. 2011. Effects of climate change on Mediterranean marine ecosystems: the case of the Catalan Sea. *Climate Research,* **50**(1), 1-29.

Carballo, R., Sanchez, M., Ramos, V., Taveira-Pinto, F. & Iglesias, G. 2014. A high resolution geospatial database for wave energy exploitation. *Energy,* **68**, 572-583.

Carrillo, C., Cidras, J., Diaz-Dorado, E. & Obando-Montano, A. F. 2014. An approach to determine the Weibull parameters for wind energy analysis: the case of Galicia (Spain). *Energies,* **7**(4), 2676-2700.

Carta, J. A., Bueno, C. & Ramirez, P. 2008a. Statistical modelling of directional wind speeds using mixtures of von Mises distributions: Case study. *Energy Conversion and Management,* **49**(5), 897-907.

Carta, J. A. & Mentado, D. 2007. A continuous bivariate model for wind power density and wind turbine energy output estimations. *Energy Conversion and Management,* **48**(2), 420-432.

Carta, J. A. & Ramirez, P. 2007. Analysis of two-component mixture Weibull statistics for estimation of wind speed distributions. *Renewable Energy,* **32**(3), 518-531.

Carta, J. A., Ramirez, P. & Bueno, C. 2008b. A joint probability density function of wind speed, and direction for wind energy analysis. *Energy Conversion and Management,* **49**(6), 1309-1320.

Carta, J. A., Ramirez, P. & Velázquez, S. 2009. A review of wind speed probability distributions used in wind energy analysis: Case studies in the Canary Islands. *Renewable & Sustainable Energy Reviews,* **13**(5), 933-955.

Carta, J. A. & Velázquez, S. 2011. A new probabilistic method to estimate the long-term wind speed characteristics at a potential wind energy conversion site. *Energy,* **36**(5), 2671-2685.

Carta, J. A., Velazquez, S. & Cabrera, P. 2013. A review of measure-correlate-predict (MCP) methods used to estimate long-term wind characteristics at a target site. *Renewable & Sustainable Energy Reviews,* **27**, 362-400.

Carvalho, D., Rocha, A., Gomez-Gesteira, M. & Santos, C. S. 2014a. Comparison of reanalyzed, analyzed, satellite-retrieved and NWP modelled winds with buoy data along the Iberian Peninsula coast. *Remote Sensing of Environment,* **152**, 480-492.

Carvalho, D., Rocha, A., Gomez-Gesteira, M. & Santos, C. S. 2014b. Offshore wind energy resource simulation forced by different reanalyses: Comparison with observed data in the Iberian Peninsula. *Applied Energy,* **134**, 57-64.

Casas-Prat, M., McInnes, K. L., Hemer, M. A. & Sierra, J. P. 2016. Future wave-driven coastal sediment transport along the Catalan coast (NW Mediterranean). *Regional Environmental Change,* **16**(6), 1739-1750.

Castillo, E. & Hadi, A. S. 1997. Fitting the generalized Pareto distribution to data. *Journal of the American Statistical Association,* **92**(440), 1609-1620.

Cebrian, A. C. & Abaurrea, J. 2006. Drought analysis based on a marked cluster Poisson model. *Journal of Hydrometeorology,* **7**(4), 713-723.

Centner, V., Massart, D. L. & de Jong, S. 1998. Inverse calibration predicts better than classical calibration. *Fresenius Journal of Analytical Chemistry,* **361**(1), 2-9.

Chang, T. P. 2011. Estimation of wind energy potential using different probability density functions. *Applied Energy,* **88**(5), 1848-1856.

Chatterjee, S. & Hadi, A. S. 1986. Influential observations, high leverage points, and outliers in linear regression. *Statistical Science,* **1**(3), 379-393.

Chellali, F., Khellaf, A., Belouchrani, A. & Khanniche, R. 2012. A comparison between wind speed distributions derived from the maximum entropy principle and Weibull distribution. Case of study; six regions of Algeria. *Renewable & Sustainable Energy Reviews,* **16**(1), 379-385.

Chen, S. X. 2000. Probability density function estimation using gamma kernels. *Annals of the Institute of Statistical Mathematics,* **52**(3), 471-480.

Cheng, C. L. & Van Ness, J. W. 2010. *Statistical regression with measurement error: Kendall's Library of Statistics 6*, Wiley.

Chowdhury, S., Zhang, J., Messac, A. & Castillo, L. 2013. Optimizing the arrangement and the selection of turbines for wind farms subject to varying wind conditions. *Renewable Energy,* **52**, 273-282.

Clementi, E., Pistoia, J., Delrosso, D., Mattia, G., Fratianni, C., Storto, A., et al. 2017. A 1/24 degree resolution Mediterranean analysis and forecast modeling system for the Copernicus Marine Environment Monitoring Service. Extended abstract to the 8th EuroGOOS Conference, Bergen.

Coco, G., Senechal, N., Rejas, A., Bryan, K. R., Capo, S., Parisot, J. P., et al. 2014. Beach response to a sequence of extreme storms. *Geomorphology,* **204**, 493-501.

Coelho, C. A. 2011. The wrapped Gamma distribution and wrapped sums and linear combinations of independent Gamma and Laplace distributions. *Journal of Statistical Theory and Practice,* **1**(1), 1-29.

Cohen, A. C. & Whitten, B. J. 1980. Estimation in the 3-Parameter Lognormal-Distribution. *Journal of the American Statistical Association,* **75**(370), 399-404.

Coles, S. 2001. *An Introduction to Statistical Modeling of Extreme Values*, Springer-Verlag London.

Coles, S. & Simiu, E. 2003. Estimating uncertainty in the extreme value analysis of data generated by a hurricane simulation model. *Journal of Engineering Mechanics-Asce,* **129**(11), 1288-1294.

Coles, S. G. & Walshaw, D. 1994. Directional modeling of extreme wind speeds. *Journal of the Royal Statistical Society Series C-Applied Statistics,* **43**(1), 139-157.

Cousineau, D. 2009. Fitting the Three-Parameter Weibull Distribution: Review and Evaluation of Existing and New Methods. *Ieee Transactions on Dielectrics and Electrical Insulation,* **16**(1), 281-288.

Cowell, P. J. & Thom, B. G. 1994. Morphodynamics of coastal evolution. *In:* Carter, R. W. G., Woodroffe, C. D. & van de Plassche, O. (eds.) *Coastal Evolution: Late Quaternary Shoreline Morphodynamics.* Cambridge: University Press.

Cran, G. W. 1988. Moment estimators for the 3-parameter Weibull distribution. *Ieee Transactions on Reliability,* **37**(4), 360-363.

Croux, C., Rousseeuw, P. J. & Hossjer, O. 1994. Generalized S-Estimators. *Journal of the American Statistical Association,* **89**(428), 1271-1281.

Cruz, J. M. B. P. & Sarmento, A. J. N. A. 2007. Sea state characterisation of the test site of an offshore wave energy plant. *Ocean Engineering,* **34**(5-6), 763-775.

Cugerone, K. & De Michele, C. 2015. Johnson SB as general functional form for raindrop size distribution. *Water Resources Research,* **51**(8), 6276-6289.

References

Cuttler, M. V. W., Hansen, J. E. & Lowe, R. J. 2020. Seasonal and interannual variability of the wave climate at a wave energy hotspot off the southwestern coast of Australia. *Renewable Energy,* **146**, 2337-2350.

Davenport, A. G. 1964. Note on the distribution of the largest value of a random function with application to gust loading. *Proc. Inst. Civ. Eng.,* **28**, 187–196.

Davison, A. C. & Smith, R. L. 1990. Models for Exceedances over High Thresholds. *Journal of the Royal Statistical Society Series B-Methodological,* **52**(3), 393-442.

de Haan, L. 1976. Sample extremes: an elementary introduction. *Statistica Neerlandica,* **30**, 161-172.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., et al. 2011. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society,* **137**(656), 553-597.

del Castillo, J. & Serra, I. 2015. Likelihood inference for generalized Pareto distribution. *Computational Statistics & Data Analysis,* **83**, 116-128.

Della-Marta, P. M., Mathis, H., Frei, C., Liniger, M. A., Kleinn, J. & Appenzeller, C. 2009. The return period of wind storms over Europe. *International Journal of Climatology,* **29**(3), 437-459.

DHI 2016. MIKE 21 Spectral Wave Module, scientific documentation. Horsholm, Denmark: Danish Hydraulics Institute.

DNV 2011. Recommended Practice DNV-RP-H103: Modelling and analysis of marine operations. Det Norske Veritas AS.

Do, K., Shin, S., Cox, D. & Yoo, J. 2018. Numerical simulation and large-scale physical modelling of coastal sand dune erosion. *Journal of Coastal Research*, Special Issue 85 - Proceedings of the 15th International Coastal Symposium: pp. 196-200.

Dodd, N., Blondeaux, P., Calvete, D., De Swart, H. E., Falques, A., Hulscher, S. J. M. H., et al. 2003. Understanding coastal morphodynamics using stability methods. *Journal of Coastal Research,* **19**(4), 849-866.

Dong, Y., Wang, J. Z., Jiang, H. & Shi, X. M. 2013. Intelligent optimized wind resource assessment and wind turbines selection in Huitengxile of Inner Mongolia, China. *Applied Energy,* **109**, 239-253.

Donoho, D. L. & Huber, P. J. 1983. The notion of breakdown point *In:* Bickel, P. J., Doksum, K. & Hodges, J. L. (eds.) *A Festschrift For Erich L. Lehmann.* Belmont (CA): Wadsworth.

Dookie, I., Rocke, S., Singh, A. & Ramlal, C. J. 2018. Evaluating wind speed probability distribution models with a novel goodness of fit metric: a Trinidad and Tobago case study. *International Journal of Energy and Environmental Engineering,* **9**(3), 323-339.

Downs, T. D. & Mardia, K. V. 2002. Circular regression. *Biometrika,* **89**(3), 683-697.

Drobinski, P., Coulais, C. & Jourdier, B. 2015. Surface wind-speed statistics modelling: alternatives to the Weibull distribution and performance evaluation. *Boundary-Layer Meteorology,* **157**(1), 97-123.

Dumouchel, W. H. 1983. Estimating the stable index-alpha in order to measure tail thickness - a critique. *Annals of Statistics,* **11**(4), 1019-1031.

Eastoe, E. F. & Tawn, J. A. 2012. Modelling the distribution of the cluster maxima of exceedances of subasymptotic thresholds. *Biometrika,* **99**(1), 43-55.

Edgeworth, F. Y. 1887. On observations relating to several quantitites. *Hermathena,* **6**(13), 279-285.

Efron, B. 1979. Bootstrap methods - another look at the jackknife. *Annals of Statistics,* **7**(1), 1-26.

Efron, B. 1987. Better bootstrap confidence intervals. *Journal of the American Statistical Association,* **82**(397), 171-185.

Efron, B. & Tibshirani, R. J. 1993. *An Introduction to the Bootstrap*, Chapman and Hall/CRC.

Eisenhart, C. 1939. The interpretation of certain regression methods and their use in biological and industrial research. *Annals of Mathematical Statistics,* **10**, 162-186.

Eldeberky, Y. & Battjes, J. A. 1996. Spectral modeling of wave breaking: Application to boussinesq equations. *Journal of Geophysical Research-Oceans,* **101**(C1), 1253-1264.

Elfrink, B., Broker, I., Deigaard, R., Hansen, E. A. & Justesen, P. 1996. Modelling of 3D sediment transport in the surf zone. *Coastal Engineering Proceedings,* **1**(25).

Embrechts, P., Klüppelberg, C. & Mikosch, T. 1997. *Modelling extremal events: for Insurance and Finance,* Berlin, Heidelberg, Springer-Verlag.

Emmanouil, G., Galanis, G., Kalogeri, C., Zodiatis, G. & Kallos, G. 2016. 10-year high resolution study of wind, sea waves and wave energy assessment in the Greek offshore areas. *Renewable Energy,* **90**, 399-419.

Engelund, F. & Fredsøe, J. 1976. Sediment transport model for straight alluvial channels. *Nordic Hydrology,* **7**(5), 293-306.

Erdem, E. & Shi, J. 2011. Comparison of bivariate distribution construction approaches for analysing wind speed and direction data. *Wind Energy,* **14**(1), 27-41.

Ewans, K. & Jonathan, P. 2008. The effect of directionality on Northern North Sea extreme wave design criteria. *Journal of Offshore Mechanics and Arctic Engineering-Transactions of the Asme,* **130**(4).

Ewans, K. C. & Jonathan, P. The effect of directionality on extreme wave design criteria. In Proceedings of the 9th International Workshop on Wave Hindcasting and Forecasting, 24-29 September 2006 Victoria, Canada. pp.

Farlie, D. J. G. 1960. The performance of some correlation coefficients for a general bivariate distribution. *Biometrika,* **47**(3-4), 307-323.

Farrugia, P. S., Borg, J. L. & Micallef, A. 2009. On the Algorithms Used to Compute the Standard Deviation of Wind Direction. *Journal of Applied Meteorology and Climatology,* **48**(10), 2144-2151.

Fawcett, L. & Walshaw, D. 2007. Improved estimation for temporally clustered extremes. *Environmetrics,* **18**(2), 173-188.

Fazeres-Ferradosa, T., Taveira-Pinto, F., Vanem, E., Reis, M. T. & das Neves, L. 2019. Asymmetric copula–based distribution models for met-ocean data in offshore wind engineering applications. *Wind Engineering,* **42**(4), 304-334.

Feng, J. & Shen, W. Z. 2015. Modelling wind for wind farm layout optimization using joint distribution of wind speed and wind direction. *Energies,* **8**(4), 3075-3092.

Fernández-Durán, J. J. 2007. Models for circular-linear and circular-circular data constructed from circular distributions based on nonnegative trigonometric sums. *Biometrics,* **63**(2), 579-585.

Ferreira, J. A. & Soares, C. G. 1999. Modelling the long-term distribution of significant wave height with the Beta and Gamma models. *Ocean Engineering,* **26**(8), 713-725.

Ferreira, M. 2018. Heuristic tools for the estimation of the extremal index: a comparison of methods. *Revstat-Statistical Journal,* **16**(1), 115-136.

Ferreira, Ó. 2005. Storm groups versus extreme single storms: predicted erosion and management consequences. *Journal of Coastal Research*, Special issue No. 42. The Sun, Earth and Moon, pp. 221-227.

Ferro, C. A. T. & Segers, J. 2003. Inference for clusters of extreme values. *Journal of the Royal Statistical Society Series B-Statistical Methodology,* **65**, 545-556.

Fisher, N. I. 1989. Smoothing a sample of circular data. *Journal of Structural Geology,* **11**(6), 775-778.

Fisher, N. I. 1993. *Statistical Analysis of Circular Data,* Cambridge, Cambridge University Press.

Fisher, N. I. 1997. Copulas. *Encyclopedia of Statistical Sciences, Update Vol. 1.* New York: John Wiley & Sons.

Fisher, N. I. & Lee, A. J. 1992. Regression models for an angular response. *Biometrics,* **48**(3), 665-677.

Fisher, R. A. 1922. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A,* **222**(594-604), 309.

Fisher, R. A. & Tippett, L. H. C. 1928. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society,* **24**, 180-190.

References

Flaounas, E., Raveh-Rubin, S., Wernli, H., Drobinski, P. & Bastin, S. 2015. The dynamical structure of intense Mediterranean cyclones. *Climate Dynamics,* **44**(9-10), 2411-2427.

Folgueras, P., Solari, S. & Losada, M. A. 2019. The selection of directional sectors for the analysis of extreme wind speed. *Natural Hazards and Earth System Sciences,* **19**(1), 221-236.

Forristall, G. Z. 2004. On the use of directional wave criteria. *Journal of Waterway Port Coastal and Ocean Engineering,* **130**(5), 272-275.

Foteinis, S. 2014. *Erosion of coastlines of Greece: assessment — responses.* PhD, Technical Univeristy of Crete.

Foteinis, S. & Synolakis, C. E. 2015. Beach erosion threatens Minoan beaches: A case study of coastal retreat in Crete. *Shore & Beach,* **83**(1), 53-62.

Fréchet, M. 1951. Sur les tableaux de corrélation dont les marges sont données. *Annales de I'Université de Lyon. Série Sci. A,* **9**, 53-77.

Fredsøe, J. 1984. Turbulent boundary-layer in wave-current motion. *Journal of Hydraulic Engineering-Asce,* **110**(8), 1103-1120.

Fuller, W. A. 1987. *Measurement error models,* New York, Wiley.

Fyrippis, I., Axaopoulos, P. J. & Panayiotou, G. 2010. Wind energy potential assessment in Naxos Island, Greece. *Applied Energy,* **87**(2), 577-586.

Günther, H. & Behrens, A. 2012. The WAM model. Validation document Version 4.5.4. Institute of Coastal Research Helmholtz-Zentrum Geesthach (HZG).

Garcia-Portugues, E. 2013. Exact risk improvement of band width selectors for kernel density estimation with directional data. *Electronic Journal of Statistics,* **7**, 1655-1685.

Gasser, T., Muller, H. G. & Mammitzsch, V. 1985. Kernels for nonparametric curve estimation. *Journal of the Royal Statistical Society Series B-Methodological,* **47**(2), 238-252.

Gillard, J. 2010. An overview of linear structural models in errors in variables regression. *Revstat-Statistical Journal,* **8**(1), 57-80.

Gnedenko, B. 1943. The limited distribution of the maximum term of a random series. *Annals of Mathematics,* **44**, 423-453.

Gong, Z., Zhang, C.-k., Zuo, C.-b. & Wu, W.-d. 2011. Sediment transport following water transfer from Yangtze River to Taihu Basin. *Water Science and Engineering,* **4**(4), 431-444.

Gould, A. L. 1969. A Regression Technique for Angular Variates. *Biometrics,* **25**(4), 683-700.

Gower, J. F. R. 1996. Intercalibration of wave and wind data from TOPEX POSEIDON and moored buoys off the west coast of Canada. *Journal of Geophysical Research-Oceans,* **101**(C2), 3817-3829.

Grabemann, I. & Weisse, R. 2008. Climate change impact on extreme wave conditions in the North Sea: an ensemble study. *Ocean Dynamics,* **58**(3-4), 199-212.

Graham, C. G. Problems with the design statistical approach. Extreme Value Analysis of Directional Wind and Wave Data, International Conference on Wave and Wind Directionality, 1981 Paris. pp. 379-401.

Greenwood, J. A., Landwehr, J. M., Matalas, N. C. & Wallis, J. R. 1979. Probability weighted moments - definition and relation to parameters of several distributions expressable in inverse form. *Water Resources Research,* **15**(5), 1049-1054.

Griffiths, G. A. 1989. A theoretically based Wakeby distribution for annual flood series. *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques,* **34**(3), 231-248.

Grimshaw, S. D. 1993. Computing maximum likelihood estimates for the Generalized Pareto distribution. *Technometrics,* **35**(2), 185-191.

Group, T. W. 1988. The WAM Model—A Third Generation Ocean Wave Prediction Model. *Journal of Physical Oceanography,* **18**(12), 1775-1810.

Guven, B. & Kotz, S. 2008. Test of independence for generalized Farlie-Gumbel-Morgenstern distributions. *Journal of Computational and Applied Mathematics,* **212**(1), 102-111.

Hall, P., Watson, G. S. & Cabrera, J. 1987. Kernel density estimation with spherical data. *Biometrika,* **74**(4), 751-762.

Hall, W. & Weller, J. 1981. Mean residual life. *In:* Csorgo, M., Dawson, D. A., Rao, J. N. K. & Saleh, A. K. M. E. (eds.) *Statistics and Related Topics* North-Holland, Amsterdam.

Hallermeier, R. J. 1980. Sand motion initiation by water waves: two asymptotes. *Journal of the Waterway Port Coastal and Ocean Division-Asce,* **106**(3), 299-318.

Hallermeier, R. J. 1981. A Profile Zonation for Seasonal Sand Beaches from Wave Climate. *Coastal Engineering,* **4**(3), 253-277.

Hallermeier, R. J. 1982. Oscillatory bedload transport - data review and simple formulation. *Continental Shelf Research,* **1**(2), 159-190.

Halperin, M. 1961. Fitting of straight lines and prediction when both variables are subject to error. *Journal of the American Statistical Association,* **56**(295), 657-669.

Halperin, M. 1970. Inverse estimation in linear regression. *Technometrics,* **12**(4), 727-736.

Hampel, F. R. 1971. A general qualitative definition of robustness. *The Annals of Mathematical Statistics,* **42**(6), 1887-1896.

Hampel, F. R. 1974. Influence curve and its role in robust estimation. *Journal of the American Statistical Association,* **69**(346), 383-393.

Hampel, F. R. 1975. Beyond location parameters: Robust concepts and methods (with discussion). *Bull. Internat. Statist. Inst.,* **46**, 375-391.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. & Stahel, W. A. 1986. *Robust statistics: the approach based on influence functions*, Wiley.

Han, Q. K., Hao, Z. L., Hu, T. & Chu, F. L. 2018. Non-parametric models for joint probabilistic distributions of wind speed and direction data. *Renewable Energy,* **126**, 1032-1042.

Härdle, W. H. 1991. *Smoothing techniques: with implementation in S*, Springer.

Harley, M. D., Turner, I. L., Kinsela, M. A., Middleton, J. H., Mumford, P. J., Splinter, K. D., et al. 2017. Extreme coastal erosion enhanced by anomalous extratropical storm wave direction. *Scientific Reports,* **7**, 6033.

Hasselmann, K. 1974. On the spectral dissipation of ocean waves due to white capping. *Boundary-Layer Meteorology,* **6**(1), 107-127.

Hasselmann, K., Barnett, T. P., Bouws, E., Carlson, H., Cartwright, D. E., Enke, K., et al. 1973. Measurements of wind-wave growth and swell decay during the Joint North Sea Wave Project (JONSWAP). *In:* JONSWAP Group (ed.) *Hydraulic Engineering Reports.* Hamburg: Deutches Hydrographisches Institut.

Hasselmann, S., Hasselmann, K., Allender, J. H. & Barnett, T. P. 1985. Computations and parameterizations of the nonlinear energy transfer in a gravity-wave spectrum. Part II: Parameterizations of the nonlinear energy transfer for application in wave models. *Journal of Physical Oceanography,* **15**(11), 1378-1391.

Haver, S. 1985. Wave climate off northern Norway. *Applied Ocean Research,* **7**(2), 85-92.

Haver, S. 1987. On the joint distribution of heights and periods of sea waves. *Ocean Engineering,* **14**(5), 359-376.

Hawkins, D. M. 1980. *Identification of outliers*, Chapman and Hall.

Heidenreich, N. B., Schindler, A. & Sperlich, S. 2013. Bandwidth selection for kernel density estimation: a review of fully automatic selectors. *Asta-Advances in Statistical Analysis,* **97**(4), 403-433.

Heritier, S., Cantoni, E., Copt, S. & Victoria-Feser, M. P. 2009. *Robust Methods in Biostatistics*, Wiley.

Hildebrandt, A., Schmidt, B. & Marx, S. 2019. Wind-wave misalignment and a combination method for direction-dependent extreme incidents. *Ocean Engineering,* **180**, 10-22.

Hiles, C. E., Beatty, S. J. & de Andres, A. 2016. Wave energy converter annual energy production uncertainty using simulations. *Journal of Marine Science and Engineering,* **4**(3).

Hipel, K. W. & McLeod, A. I. 1994. *Time series modelling of water resources and environmental systems,* Amsterdam, Elsevier.

Hirose, H. 1997. Maximum likelihood parameter estimation in the three-parameter log-normal distribution using the continuation method. *Computational Statistics & Data Analysis,* **24**(2), 139-152.

Hirukawa, M. 2018. *Asymmetric kernel smoothing,* Singapore, Springer.

Hirukawa, M. & Sakudo, M. 2014. Nonnegative bias reduction methods for density estimation using asymmetric kernels. *Computational Statistics & Data Analysis,* **75**, 112-123.

Hoaglin, D. C. & Welsch, R. E. 1978. Hat matrix in regression and anova. *American Statistician,* **32**(1), 17-22.

Hodges, J. L. & Lehmann, E. L. 1963. Estimates of location based on rank tests. *Ann. Math. Statist.,* **34**(2), 598-611.

Hogg, R. V. & Klugman, S. A. 1984. *Loss distributions,* New York, Wiley.

Holthuijsen, L. H., Booij, N. & Ris, R. C. A spectral wave model for the coastal zone. 2nd International Symposium on Ocean Wave Measurement and Analysis, July 25-28, 1993 1993 New Orleans, Louisiana. New York, pp. 630-641.

Horn, J. T., Krokstad, J. R. & Amdahl, J. 2018. Long-term fatigue damage sensitivity to wave directionality in extra-large monopile foundations. *Proceedings of the Institution of Mechanical Engineers Part M-Journal of Engineering for the Maritime Environment,* **232**(1), 37-49.

Hoshi, K., Stedinger, J. R. & Burges, S. J. 1984. Estimation of log-normal quantiles - Monte Carlo results and first-order approximations. *Journal of Hydrology,* **71**(1-2), 1-30.

Hosking, J. R. M. 1994. The four-parameter Kappa distribution. *Ibm Journal of Research and Development,* **38**(3), 251-258.

Hosking, J. R. M. & Wallis, J. R. 1987. Parameter and quantile estimation for the Generalized Pareto distribution. *Technometrics,* **29**(3), 339-349.

Hosking, J. R. M., Wallis, J. R. & Wood, E. F. 1985. Estimation of the Generalized Extreme Value distribution by the method of probability weighted moments. *Technometrics,* **27**(3), 251-261.

Hossjer, O. 1992. On the optimality of S-estimators. *Statistics & Probability Letters,* **14**(5), 413-419.

Hu, B., Li, Y., Yang, H. & Wang, H. 2016. Wind speed model based on kernel density estimation and its application in reliability assessment of generating systems. *Journal of Modern Power Systems and Clean Energy,* **5**(2), 220-227.

Huang, J. L. & McElroy, M. B. 2015. A 32-year perspective on the origin of wind energy in a warming climate. *Renewable Energy,* **77**, 482-492.

Huber, P. J. 1964. Robust estimation of location parameter. *Annals of Mathematical Statistics,* **35**(1), 73-101.

Huber, P. J. 1981. *Robust statistics,* New York, Wiley.

Huber, P. J. & Ronchetti, E. M. 2009. *Robust statistics*, Wiley.

Hussin, A. G., Fieller, N. R. J. & Stillman, E. C. 2004. Linear regression model for circular variables with application to directional data. *Journal of Applied Science & Technology,* **9**(1&2), 1-6.

Hyndman, R. J. & Koehler, A. B. 2006. Another look at measures of forecast accuracy. *International Journal of Forecasting,* **22**(4), 679-688.

Igarashi, G. 2016. Weighted log-normal kernel density estimation. *Communications in Statistics-Theory and Methods,* **45**(22), 6670-6687.

Ikeda, S., Osawa, K. & Akamatsu, Y. 2009. Sediment and nutrients transport in watershed and their impact on coastal environment. *Proceedings of the Japan Academy Series B-Physical and Biological Sciences,* **85**(9), 374-390.

Imon, A. H. M. R. 2005. Identifying multiple influential observations in linear regression. *Journal of Applied Statistics,* **32**(9), 929-946.

Jaeckel, L. A. 1972. Estimating regression coefficients by minimizing dispersion of residuals. *Annals of Mathematical Statistics,* **43**(5), 1449-1458.

Jammalamadaka, R. S. & SenGupta, A. (eds.) 2001. *Topics in circular statistics,* Singapore: World Scientific Publishing Co. Pte. Ltd.

Jammalamadaka, S. R. & Sarma, Y. R. 1993. Circular regression. *In:* Matusita, K. (ed.) *Statistical Sciences and Data Analysis.* Utrecht: VSP.

Janssen, P. A. E. M. 1989. Wave-induced stress and the drag of air flow over sea waves. *Journal of Physical Oceanography,* **19**(6), 745-754.

Janssen, P. A. E. M. 1991. Quasi-linear theory of wind-wave generation applied to wave forecasting. *Journal of Physical Oceanography,* **21**(11), 1631-1642.

Janssen, P. A. E. M. Consequences of the effect of surface gravity waves on the mean air flow. 1992 Berlin, Heidelberg. Springer Berlin Heidelberg, pp. 193-198.

Janssen, P. A. E. M., Lionello, P. & Zambresky, L. 1989. On the Interaction of Wind and Waves. *Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences,* **329**(1604), 289-301.

Jenkinson, A. F. 1955. The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society,* **81**(348), 158-171.

Jeon, J. & Taylor, J. W. 2012. Using conditional kernel density estimation for wind power density forecasting. *Journal of the American Statistical Association,* **107**(497), 66-79.

Jeon, Y. & Kim, J. H. T. 2013. A gamma kernel density estimation for insurance loss data. *Insurance Mathematics & Economics,* **53**(3), 569-579.

Jia, J. B. 2011. Wind and structural modelling for an accurate fatigue life assessment of tubular structures. *Engineering Structures,* **33**(2), 477-491.

Jiang, C., Chen, J., Yao, Y., Liu, J. & Deng, Y. 2015. Study on threshold motion of sediment and bedload transport by tsunami waves. *Ocean Engineering,* **100**, 97-106.

Jin, X. & Kawczak, J. 2003. Birnbaum-Saunders and Lognormal kernel estimators for modelling durations in high frequency financial data. *Annals of Economics and Finance,* **4**, 103-124.

Jing-Jing, X., Fei, H., Zi-Niu, X. & Xue-Ling, C. 2014. Bias correction in wind direction forecasting using the circular-circular regression method. *Atmospheric and Oceanic Science Letters,* **7**(2), 87-91.

Johnson, H. K. & Kofoed-Hansen, H. 2000. Influence of bottom friction on sea surface roughness and its impact on shallow water wind wave modeling. *Journal of Physical Oceanography,* **30**(7), 1743-1756.

Johnson, N. L. 1949. Systems of frequency curves generated by methods of translation. *Biometrika,* **36**(1-2), 149-176.

Johnson, N. L., Kotz, S. & Balakrishnan, N. 1995. *Continuous univariate distributions,* New York, Wiley & Sons.

Johnson, R. A. & Wehrly, T. E. 1978. Some angular-linear distributions and related regression models. *Journal of the American Statistical Association,* **73**(363), 602-606.

Jonathan, P. & Ewans, K. 2007. The effect of directionality on extreme wave design criteria. *Ocean Engineering,* **34**(14-15), 1977-1994.

Jonathan, P., Ewans, K. & Forristall, G. 2008. Statistical estimation of extreme ocean environments: The requirement for modelling directionality and other covariate effects. *Ocean Engineering,* **35**(11-12), 1211-1225.

Jourdier, B. & Drobinski, P. 2017. Errors in wind resource and energy yield assessments based on the Weibull distribution. *Annales Geophysicae,* **35**(3), 691-700.

Jung, C. & Schindler, D. 2019. Wind speed distribution selection – A review of recent development and progress. *Renewable and Sustainable Energy Reviews,* **114**, 109290.

Kamphuis, J. W. 1975. Friction factor under oscillatory waves. *Journal of the Waterways Harbors and Coastal Engineering Division-ASCE,* **101**(2), 135-144.

Kannan, N., Keating, J. P. & Mason, R. L. 2007. A comparison of classical and inverse estimators in the calibration problem. *Communications in Statistics-Theory and Methods,* **36**(1-4), 83-95.

Kapelonis, Z. G., Gavriliadis, P. N. & Athanassoulis, G. A. 2015. Extreme value analysis of dynamical wave climate projections in the Mediterranean Sea. *Procedia Computer Science,* **66**, 210-219.

Karathanasi, F., Belibassakis, K. & Anagnostou, C. 2017. Simulation of wave field and sediment transport at the Sitia bay. *7th National Conference on Management and Improvement of Coastal Zones.* 20-22 November, Athens.

Karathanasi, F., Soukissian, T. & Sifnioti, D. 2015. Offshore wave potential of the Mediterranean Sea. *Renewable Energy and Power Quality Journal,* **1**(13), 461-465.

References

Karathanasi, F. E., Soukissian, T. H. & Axaopoulos, P. G. 2016. Calibration of Wind Directions in the Mediterranean Sea. *The 26th International Ocean and Polar Engineering Conference.* Rhodes, Greece: International Society of Offshore and Polar Engineers.

Karimirad, M. & Michailides, C. 2016. V-shaped semisubmersible offshore wind turbine subjected to misaligned wave and wind. *Journal of Renewable and Sustainable Energy,* **8**(2).

Karunamuni, R. J. & Alberts, T. 2005. On boundary correction in kernel density estimation. *Statistical Methodology,* (2), 191-212.

Karunamuni, R. J. & Zhang, S. 2008. Some improvements on a boundary corrected kernel density estimator. *Statistics & Probability Letters,* **78**(5), 499-507.

Karunarathna, H., Horrillo-Caraballo, J. M. & Reeve, D. E. 2012. Prediction of cross-shore beach profile evolution using a diffusion type model. *Continental Shelf Research,* **48**, 157-166.

Kato, S. & Jones, M. C. 2010. A Family of Distributions on the Circle With Links to, and Applications Arising From, Mobius Transformation. *Journal of the American Statistical Association,* **105**(489), 249-262.

Kato, S. & Jones, M. C. 2013. An extended family of circular distributions related to wrapped Cauchy distributions via Brownian motion. *Bernoulli,* **19**(1), 154-171.

Kato, S., Shimizu, K. & Shieh, G. S. 2008. A circular-circular regression model. *Statistica Sinica,* **18**(2), 633-645.

Katz, R. W., Parlange, M. B. & Naveau, P. 2002. Statistics of extremes in hydrology. *Advances in Water Resources,* **25**(8-12), 1287-1304.

Kendall, M. G. & Stuart, A. 1973. *The Advanced Theory of Statistics,* London, Charles Griffin and Co. Ltd.

Kent, J. T. & Tyler, D. E. 1988. Maximum likelihood estimation for the wrapped Cauchy distribution. *Journal of Applied Statistics,* **15**, 247-254.

Kim, S. & Sengupta, A. 2015. Inverse circular-linear/linear-circular regression. *Communications in Statistics-Theory and Methods,* **44**(22), 4772-4782.

Kirby, J. T. & Dalrymple, R. A. 1986. An approximate model for nonlinear dispersion in monochromatic wave propagation models. *Coastal Engineering,* **9**(6), 545-561.

Kjeldsen, T. R., Ahn, H. & Prosdocimi, I. 2017. On the use of a four-parameter kappa distribution in regional frequency analysis. *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques,* **62**(9), 1354-1363.

Klepper, S. & Leamer, E. E. 1984. Consistent Sets of Estimates for Regressions with Errors in All Variables. *Econometrica,* **52**(1), 163-183.

Klonaris, G. T., Memos, C. D., Dronen, N. K. & Deigaard, R. 2018. Simulating 2DH coastal morphodynamics with a Boussinesq-type model. *Coastal Engineering Journal,* **60**(2), 159-179.

Kluyver, J. C. 1906. A local probability theorem. *Ned. Akad. Wet. Proc.,* **A8**, 341-350.

Koenker, R. & Bassett, G. 1978. Regression Quantiles. *Econometrica,* **46**(1), 33-50.

Koletsis, I., Kotroni, V., Lagouvardos, K. & Soukissian, T. 2016. Assessment of offshore wind speed and power potential over the Mediterranean and the Black Seas under future climate changes. *Renewable & Sustainable Energy Reviews,* **60**, 234-245.

Kollu, R., Rayapudi, S. R., Narasimham, S. & Pakkurthi, K. M. 2012. Mixture probability distribution functions to model wind speed distributions. *International Journal of Energy and Environmental Engineering,* **3**(1), 27.

Komen, G. J., Cavaleri, L., Donelan, M., Hasselmann, K., Hasselmann, S. & Janssen, P. A. E. M. 1994. *Dynamics and Modelling of Ocean Waves*, Cambridge University Press.

Kontoyiannis, H. 2010. Observations on the circulation of the Saronikos Gulf: A Mediterranean embayment sea border of Athens, Greece. *Journal of Geophysical Research-Oceans,* **115**.

Korres, G., Nitti, K., Perivoliotis, L., Tsiaras, K., Papadopoulos, A., Triantafyllou, G., et al. 2010. Forecasting the Aegean Sea hydrodynamics within the POSEIDON-II operational system. *Journal of Operational Oceanography,* **3**(1), 37-49.

Kotz, S., Balakrishnan, N. & Johnson, N. L. 2000. *Continuous multivariate distributions, Volume 1: Models and Applications,* New York, John Wiley & Sons.

Krutchkoff, R. G. 1967. Classical and inverse regression methods of calibration. *Technometrics,* **9**(3), 425-+.

Lackner, M. A., Rogers, A. L. & Manwell, J. F. 2008. Uncertainty analysis in MCP-Based wind resource assessment and energy production estimation. *Journal of Solar Energy Engineering-Transactions of the Asme,* **130**(3).

Lagouvardos, K., Kotroni, V. & Defer, E. 2007. The 21-22 January 2004 explosive cyclogenesis over the Aegean Sea: Observations and model analysis. *Quarterly Journal of the Royal Meteorological Society,* **133**(627), 1519-1531.

Landwehr, J. M., Matalas, N. C. & Wallis, J. R. 1979a. Estimation of Parameters and Quantiles of Wakeby Distributions. 1. Known Lower Bounds. *Water Resources Research,* **15**(6), 1361-1372.

Landwehr, J. M., Matalas, N. C. & Wallis, J. R. 1979b. Estimation of Parameters and Quantiles of Wakeby Distributions. 2. Unknown Lower Bounds. *Water Resources Research,* **15**(6), 1373-1379.

Langousis, A., Mamalakis, A., Puliga, M. & Deidda, R. 2016. Threshold detection for the generalized Pareto distribution: Review of representative methods and application to the NOAA NCDC daily rainfall database. *Water Resources Research,* **52**(4), 2659-2681.

Larsén, X. G., Kalogeri, C., Galanis, G. & Kallos, G. 2015. A statistical methodology for the estimation of extreme wave conditions for offshore renewable applications. *Renewable Energy,* **80**, 205-218.

Laycock, P. J. 1975. Optimal Design - Regression Models for Directions. *Biometrika,* **62**(2), 305-311.

Leadbetter, M. R. 1974. Extreme Values in Stationary Sequences. *Zeitschrift Fur Wahrscheinlichkeitstheorie Und Verwandte Gebiete,* **28**(4), 289-303.

Ledford, A. W. & Tawn, J. A. 2003. Diagnostics for dependence within time series extremes. *Journal of the Royal Statistical Society Series B-Statistical Methodology,* **65**, 521-543.

Lerma, A. N., Bulteau, T., Lecacheux, S. & Idier, D. 2015. Spatial variability of extreme wave height along the Atlantic and channel French coast. *Ocean Engineering,* **97**, 175-185.

Leys, C., Klein, O., Dominicy, Y. & Ley, C. 2018. Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. *Journal of Experimental Social Psychology,* **74**, 150-156.

Li, L. L. & Huang, Z. H. 2013. Modeling the change of beach profile under tsunami waves: a comparison of selected sediment transport models. *Journal of Earthquake and Tsunami,* **7**(1).

Lindley, D. V. 1947. Regression lines and the linear functional relationship. *Supplement to the Journal of the Royal Statistical Society,* **9**(2), 218-244.

Lizuma, L., Avotniece, Z., Rupainis, S. & Teilans, A. 2013. Assessment of the present and future offshore wind power potential: a case study in a target territory of the Baltic Sea near the Latvian coast. *Scientific World Journal*.

Loader, C. R. 1999. Bandwidth selection: Classical or plug-in? *Annals of Statistics,* **27**(2), 415-438.

Long, D. & Krzysztofowicz, R. 1992. Farlie-Gumbel-Morgenstern bivariate densities: Are they applicable in hydrology? *Stochastic Hydrology and Hydraulics,* **6**(1), 47-54.

Lu, Y. J., Li, S. Q., Zuo, L. Q., Liu, H. X. & Roelvink, J. A. 2015. Advances in sediment transport under combined action of waves and currents. *International Journal of Sediment Research,* **30**(4), 351-360.

Lucas, C. & Soares, C. G. 2015. Bivariate distributions of significant wave height and mean wave period of combined sea states. *Ocean Engineering,* **106**, 341-353.

Lund, U. 1999. Least circular distance regression for directional data. *Journal of Applied Statistics,* **26**(6), 723-733.

Luo, S. L., Liu, Y. H., Jin, R. F., Zhang, J. P. & Wei, W. 2016. A guide to coastal management: Benefits and lessons learned of beach nourishment practices in China over the past two decades. *Ocean & Coastal Management,* **134**, 207-215.

Lyu, G., Zhang, H. & Li, J. 2019. Effects of incident wind/wave directions on dynamic response of a SPAR-type floating offshore wind turbine system. *Acta Mechanica Sinica*, 1-10.

Madansky, A. 1959. The fitting of straight lines when both variables are subject to error. *Journal of the American Statistical Association,* **54**(285), 173-205.

Madsen, O. S. 1994. Spectral wave-current bottom boundary layer flows. *in: Coastal Engineering Proceedings, 24th International Conference Coastal Engineering Research Council.* Kobe, Japan: American Society of Civil Engineers.

Madsen, P. A., Murray, R. & Sørensen, O. R. 1991. A new form of the Boussinesq equations with improved linear dispersion characteristics. *Coastal Engineering,* **15**(4), 371-388.

Mallows, C. 1975. On some topics in robustness. Technical memorandum. Murray Hill, NJ: Bell Telephone Laboratories.

Marchant, C., Bertin, K., Leiva, V. & Saulo, H. 2013. Generalized Birnbaum-Saunders kernel density estimators and an analysis of financial data. *Computational Statistics & Data Analysis,* **63**, 1-15.

Mardia, K. V. 1967. A non-parametric test for the bivariate two-sample location problem. *Journal of the Royal Statistical Society Series B-Statistical Methodology,* **29**(2), 320-342.

Mardia, K. V. 1970a. *Families of bivariate distributions,* London, Griffin.

Mardia, K. V. 1970b. A translation family of bivariate distributions and Fréchet's bounds. *Sankhyā: The Indian Journal of Statistics, Series A* **32**, 119-121.

Mardia, K. V. 1972. Statistics of Directional Data. *Journal of the Royal Statistical Society Series B-Statistical Methodology,* **37**(3), 349-393.

Mardia, K. V. 1976. Linear-Circular Correlation-Coefficients and Rhythmometry. *Biometrika,* **63**(2), 403-405.

Mardia, K. V. & Jupp, P. E. 2009. *Directional Statistics*, Wiley.

Mardia, K. V. & Spurr, B. D. 1973. Multisample tests for multimodal and axial circular populations. *Journal of the Royal Statistical Society Series B-Methodological,* **35**(3), 422-436.

Marine Information Service 2016. EMODnet Digital Bathymetry (DTM). Marine Information Service.

Marković, D., Jukić, D. & Benšić, M. 2009. Nonlinear weighted least squares estimation of a three-parameter Weibull density with a nonparametric start. *Journal of Computational and Applied Mathematics,* **228**(1), 304-312.

Maronna, R. A., Martin, D. R. & Yohai, V. J. 2006. *Robust Statistics: Theory and Methods*, Wiley.

Marron, J. S. & Ruppert, D. 1994. Transformations to reduce boundary bias in kernel density estimation. *Journal of the Royal Statistical Society Series B-Methodological,* **56**(4), 653-671.

Masseran, N., Razali, A. M., Ibrahim, K. & Latif, M. T. 2013. Fitting a mixture of von Mises distributions in order to model data on wind direction in Peninsular Malaysia. *Energy Conversion and Management,* **72**, 94-102.

Mathisen, J. & Bitner-Gregersen, E. 1990. Joint distributions for significant wave height and wave zero-up-crossing period. *Applied Ocean Research,* **12**(2), 93-103.

McAssey, M. P. 2013. An empirical goodness-of-fit test for multivariate distributions. *Journal of Applied Statistics,* **40**(5), 1120-1131.

Menendez, M., García-Díez, M., Fita, L., Fernández, J., Méndez, F. J. & Gutiérrez, J. M. 2014. High-resolution sea wind hindcasts over the Mediterranean area. *Climate Dynamics,* **42**(7-8), 1857-1872.

Menjoge, R. S. & Welsch, R. E. 2010. A diagnostic method for simultaneous feature selection and outlier identification in linear regression. *Computational Statistics & Data Analysis,* **54**(12), 3181-3193.

Moarefdoost, M. M., Snyder, L. V. & Alnajjab, B. 2017. Layouts for ocean wave energy farms: Models, properties, and optimization. *Omega-International Journal of Management Science,* **66**, 185-194.

Mooney, J. A., Helms, P. J. & Jolliffe, I. T. 2003. Fitting mixtures of von Mises distributions: a case study involving sudden infant death syndrome. *Computational Statistics & Data Analysis,* **41**(3-4), 505-513.

Morgan, E. C., Lackner, M., Vogel, R. M. & Baise, L. G. 2011. Probability distributions for offshore wind speeds. *Energy Conversion and Management,* **52**(1), 15-26.

Morgenstern, D. 1956. Einfache Beispiele zweidimensionaler Verteilungen. *Mitteilungsblatt für Mathematishe Statistik,* **8**, 234-235.

Moriarty, W. W. & Templeton, J. I. 1983. On the estimation of extreme wind gusts by direction sector. *Journal of Wind Engineering and Industrial Aerodynamics,* **13**(1-3), 127-138.

Morrissey, M. L., Cook, W. E. & Greene, J. S. 2010. An improved method for estimating the wind power density distribution function. *Journal of Atmospheric and Oceanic Technology,* **27**(7), 1153-1164.

Mortlock, T. R., Goodwin, I. D., McAneney, J. K. & Roche, K. 2017. The June 2016 Australian east coast low: importance of wave direction for coastal erosion assessment. *Water,* **9**(2).

Morton, I. D., Bowers, J. & Mould, G. 1997. Estimating return period wave heights and wind speeds using a seasonal point process model. *Coastal Engineering,* **31**(1-4), 305-326.

Muller, H. G. 1991. Smooth optimum kernel estimators near endpoints. *Biometrika,* **78**(3), 521-530.

Muraleedharan, G., Lucas, C., Martins, D., Soares, C. G. & Kurup, P. G. 2015. On the distribution of significant wave height and associated peak periods. *Coastal Engineering,* **103**, 42-51.

Muraleedharan, G., Sinha, M., Rao, A. D., Nair, N. & Kurup, P. G. 2009. Estimation of wave period statistics using numerical coastal wave model. *Natural Hazards,* **49**(2), 165-186.

Murshed, M. S., Seo, Y. A. & Park, J. S. 2014. LH-moment estimation of a four parameter kappa distribution with hydrologic applications. *Stochastic Environmental Research and Risk Assessment,* **28**(2), 253-262.

Murthy, D. N. P., Xie, M. & Jiang, R. 2004. *Weibull Models,* New York, John Wiley.

Nelsen, R. B. 2006. *An introduction to copulas,* New York, Springer.

Nerzic, R. & Prevosto, M. Modelling of wind and wave joint occurrence probability and persistence duration from satellite observation data In: 10th International Offshore and Polar Engineering Conference, May 28-June 2 2000 Seattle, USA. pp. 154-158.

Neumann, B., Vafeidis, A. T., Zimmermann, J. & Nicholls, R. J. 2015. Future Coastal Population Growth and Exposure to Sea-Level Rise and Coastal Flooding - A Global Assessment. *Plos One,* **10**(6), e0118571.

Ng, C. & Ran, L. 2016. *Offshore wind farms: technologies design and operation*, Woodhead Publishing.

Nielsen, P. 1992. *Coastal bottom boundary layers and sediment transport,* Singapore, World Scientific.

Nielsen, P. & Callaghan, D. P. 2003. Shear stress and sediment transport calculations for sheet flow under waves. *Coastal Engineering,* **47**(3), 347-354.

Nurunnabi, A. A. M., Hadi, A. S. & Imon, A. H. M. R. 2014. Procedures for the identification of multiple influential observations in linear regression. *Journal of Applied Statistics,* **41**(6), 1315-1331.

Nurunnabi, A. A. M., Nasser, M. & Imon, A. H. M. R. 2016. Identification and classification of multiple outliers, high leverage points and influential observations in linear regression. *Journal of Applied Statistics,* **43**(3), 509-525.

Ochi, M. K. New approach for estimating the severest sea state from statistical data. In: 23rd International Conference on Coastal Engineering, 4-9 October 1992 Venice, Italy. pp.

Oliveira, M., Crujeiras, R. M. & Rodriguez-Casal, A. 2012. A plug-in rule for bandwidth selection in circular density estimation. *Computational Statistics & Data Analysis,* **56**(12), 3898-3908.

References

Ortiz, A. C. & Ashton, A. D. 2016. Exploring shoreface dynamics and a mechanistic explanation for a morphodynamic depth of closure. *Journal of Geophysical Research-Earth Surface,* **121**(2), 442-464.

Osborne, C. 1991. Statistical calibration - a review. *International Statistical Review,* **59**(3), 309-336.

Ouarda, T. B. M. J. & Charron, C. 2018. On the mixture of wind speed distribution in a Nordic region. *Energy Conversion and Management,* **174**, 33-44.

Ouarda, T. B. M. J., Charron, C. & Chebana, F. 2016. Review of criteria for the selection of probability distributions for wind speed data and introduction of the moment and L-moment ratio diagram methods, with a case study. *Energy Conversion and Management,* **124**, 247-265.

Ouarda, T. B. M. J., Charron, C., Shin, J. Y., Marpu, P. R., Al-Mandoos, A. H., Al-Tamimi, M. H., et al. 2015. Probability distributions of wind speed in the UAE. *Energy Conversion and Management,* **93**, 414-434.

Öztekin, T. 2011. Estimation of the parameters of Wakeby distribution by a numerical least squares method and applying it to the annual peak flows of Turkish rivers. *Water Resources Management,* **25**(5), 1299-1313.

Papanicolaou, A. N., Elhakeem, M., Krallis, G., Prakash, S. & Edinger, J. 2008. Sediment transport modeling review - Current and future developments. *Journal of Hydraulic Engineering,* **134**(1), 1-14.

Park, J. S. & Jung, H. S. 2002. Modelling Korean extreme rainfall using a Kappa distribution and maximum likelihood estimate. *Theoretical and Applied Climatology,* **72**(1-2), 55-64.

Pearson, K. 1905a. The problem of the random walk. *Nature,* **72**, 342-342.

Pearson, K. 1905b. The problem of the random walk. *Nature,* **72**, 294-294.

Pearson, K. 1906. A Mathematical Theory of Random Migration. *In:* Pearson, K. & Nettleship , E. (eds.) *Draper's company research memoirs.*

Perera, L. P. & Soares, C. G. 2017. Weather routing and safe ship handling in the future of shipping. *Ocean Engineering,* **130**, 684-695.

Pewsey, A., Lewis, T. & Jones, M. C. 2007. The wrapped t family of circular distributions. *Australian & New Zealand Journal of Statistics,* **49**(1), 79-91.

Pewsey, A., Neuhäuser, M. & Ruxton, G. D. 2013. *Circular statistics in R*, OUP Oxford.

Phien, H. N. & Jivajirajah, T. 1984. Fitting the SB Curve by the method of maximum likelihood. *Journal of Hydrology,* **67**(1-4), 67-75.

Philippe, M., Babarit, A. & Ferrant, P. 2013. Modes of response of an offshore wind turbine with directional wind and waves. *Renewable Energy,* **49**, 151-155.

Pickands, J. 1975. Statistical inference using extreme order statistics. *Annals of Statistics,* **3**(1), 119-131.

Plackett, R. L. 1965. A class of bivariate distributions. *Journal of the American Statistical Association,* **60**(310), 516-522.

Polsen, O. & Taylor, C. C. 2015. Parametric circular–circular regression and diagnostic analysis. *In:* Dryden, I. L. & Kent, J. T. (eds.) *Geometry Driven Statistics.*

Porté-Agel, F., Wu, Y. T. & Chen, C. H. 2013. A numerical study of the effects of wind direction on turbine wakes and power losses in a large wind farm. *Energies,* **6**(10), 5297-5313.

Poupkou, A., Zanis, P., Nastos, P., Papanastasiou, D., Melas, D., Tourpali, K., et al. 2011. Present climate trend analysis of the Etesian winds in the Aegean Sea. *Theoretical and Applied Climatology,* **106**(3-4), 459-472.

Qin, X., Zhang, J. & Yan, X. 2010. A new circular distribution and its application to wind data. *Journal of Mathematics Research,* **2**(3), 12-17.

Qin, X., Zhang, J. S. & Yan, X. D. 2012. A finite mixture three-parameter Weibull model for the analysis of wind speed data. *Communications in Statistics-Theory and Methods,* **41**(12), 2160-2171.

Qu, X. L. & Shi, J. 2010. Bivariate modeling of wind speed and air density distribution for long-term wind energy estimation. *International Journal of Green Energy,* **7**(1), 21-37.

Ramšak, V., Malačič, V., Ličer, M., Kotnik, J., Horvat, M. & Žagar, D. 2013. High-resolution pollutant dispersion modelling in contaminated coastal sites. *Environmental Research,* **125**, 103-112.

Rao, A. R. & Hamed, K. H. 2000. *Flood frequency analysis.,* New York, CRC Press.

Rayleigh 1905. The problem of the random walk. *Nature,* **72**, 318-318.

Rayleigh 1919. On the problem of random vibrations, and of random flights in one, two, or three dimensions. *Philosophical Magazine,* **37**(220), 321-347.

Rayleigh, L. 1880. XII. On the resultant of a large number of vibrations of the same pitch and of arbitrary phase. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science,* **10**(60), 73-78.

Reiss, R.-D. & Thomas, M. 2007. *Statistical analysis of extreme ealues, with application to insurance, finance, hydrology and other fields,* 3rd ed., Switzerland, Birkhäuser Verlag.

Revfeim, K. J. A. 1983. On the analysis of extreme rainfalls. *Journal of Hydrology,* **62**(1-4), 107-117.

Riggs, D. S., Guarnieri, J. A. & Addelman, S. 1978. Fitting straight lines when both variables are subject to error. *Life Sciences,* **22**(13-1), 1305-1360.

Rivest, L.-P. 1997. A decentred predictor for circuclar-circular regression. *Biometrika,* **84**(3), 717-726.

Robinson, M. E. & Tawn, J. A. 1997. Statistics for extreme sea currents. *Journal of the Royal Statistical Society Series C-Applied Statistics,* **46**(2), 183-205.

Rocha, P. A. C., de Sousa, R. C., de Andrade, C. F. & da Silva, M. E. V. 2012. Comparison of seven numerical methods for determining Weibull parameters for wind energy generation in the northeast region of Brazil. *Applied Energy,* **89**(1), 395-400.

Rodriguez, O., del Rio, J. A., Jaramillo, O. A. & Martinez, M. 2015. Wind power error estimation in resource assessments. *Plos One,* **10**(5).

Roe, P. L. 1981. Approximate Riemann solvers, parameter vectors, and difference schemes. *Journal of Computational Physics,* **43**(2), 357-372.

Rosenblatt, M. 1956. Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics,* **27**(3), 832-837.

Rousseeuw, P. & Yohai, V. Robust regression by means of S-estimators. 1984 New York, NY. Springer US, pp. 256-272.

Rousseeuw, P. J. 1984. Least median of squares regression. *Journal of the American Statistical Association,* **79**(388), 871-880.

Rousseeuw, P. J. & Leroy, A. M. 1987. *Robust regression and outlier detection*, Wiley.

Rueda, C., Fernandez, M. A., Barragan, S., Mardia, K. V. & Peddada, S. D. 2016. Circular piecewise regression with applications to cell-cycle data. *Biometrics,* **72**(4), 1266-1274.

Ryan, T. P. 1997. *Modern regression methods*, John Wiley & Sons.

Samayam, S., Laface, V., Annamalaisamy, S. S., Arena, F., Vallam, S. & Gavrilovich, P. V. 2017. Assessment of reliability of extreme wave height prediction models. *Natural Hazards and Earth System Sciences,* **17**(3), 409-421.

Sanchez-Arcilla, A., Aguar, J. G., Egozcue, J. J., Ortego, M. I., Galiatsatou, P. & Prinos, P. 2008. Extremes from scarce data: The role of Bayesian and scaling techniques in reducing uncertainty. *Journal of Hydraulic Research,* **46**(2), 224-234.

Santos, V. M., Haigh, I. D. & Wahl, T. 2017. Spatial and temporal clustering analysis of extreme wave events around the UK coastline. *Journal of Marine Science and Engineering,* **5**(3).

Sarkar, A., Gugliani, G. & Deep, S. 2017. Weibull model for wind speed data analysis of different locations in India. *Ksce Journal of Civil Engineering,* **21**(7), 2764-2776.

Sarma, R., Rao, A. V. D. & Girija, S. V. S. 2011. On characteristic functions of the wrapped lognormal and the wrapped Weibull distributions. *Journal of Statistical Computation and Simulation,* **81**(5), 579-589.

Scaillet, O. 2004. Density estimation using inverse and reciprocal inverse Gaussian kernels. *Journal of Nonparametric Statistics,* **16**(1-2), 217-226.

Scarrott, C. & MacDonald, A. 2012. A review of extreme value threshold estimation and uncertainty quantification. *Revstat-Statistical Journal,* **10**(1), 33-60.

Schmidt, K. M., Swart, S., Reason, C. & Nicholson, S. A. 2017. Evaluation of satellite and reanalysis wind products with in situ wave glider wind observations in the Southern Ocean. *Journal of Atmospheric and Oceanic Technology,* **34**(12), 2551-2568.

Schuster, E. F. 1985. Incorporating support constraints into nonparametric estimators of densities. *Communications in Statistics-Theory and Methods,* **14**(5), 1123-1136.

Schwarz, G. 1978. Estimating dimension of a model. *Annals of Statistics,* **6**(2), 461-464.

Scolforo, J. R. S., Tabai, F. C. V., de Macedo, R. L. G., Acerbi, F. W. & de Assis, A. L. 2003. S-B distribution's accuracy to represent the diameter distribution of Pinus taeda, through five fitting methods. *Forest Ecology and Management,* **175**(1-3), 489-496.

SenGupta, A. & Kim, S. 2016. Statistical inference for homologous gene pairs between two circular genomes: a new circular–circular regression model. *Statistical Methods & Applications,* **25**(3), 421-432.

SenGupta, A., Kim, S. & Arnold, B. C. 2013. Inverse circular-circular regression. *Journal of Multivariate Analysis,* **119**, 200-208.

SenGupta, A. & Ugwuowo, F. I. 2006. Asymmetric circular-linear multivariate regression models with applications to environmental data. *Environmental and Ecological Statistics,* **13**(3), 299-309.

Serafim, M. B., Siegle, E., Corsi, A. C. & Bonetti, J. 2019. Coastal vulnerability to wave impacts using a multi-criteria index: Santa Catarina (Brazil). *Journal of Environmental Management,* **230**, 21-32.

Shields, I. A. 1936. *Anwendung der Aehnlichkeitsmechanik and der Turbulenzforschung auf die Geschiebebewegung,* Tech. Univ. Berlin, Berlin, Eigenverl. der Preußischen Versuchsanst. für Wasserbau und Schiff.

Shin, J. Y., Ouarda, T. B. M. J. & Lee, T. 2016. Heterogeneous mixture distributions for modeling wind speed, application to the UAE. *Renewable Energy,* **91**, 40-52.

Sierra, J. P., Genius, A., Lionello, P., Mestres, M., Mosso, C. & Marzo, L. 2017. Modelling the impact of climate change on harbour operability: The Barcelona port case study. *Ocean Engineering,* **141**, 64-78.

Silverman, B. W. 1986. *Density estimation for statistics and data analysis*, Taylor & Francis.

Simiu, E., Hendrickson, E. M., Nolan, W. A., Olkin, I. & Spiegelman, C. H. 1985. Multivariate distributions of directional wind speeds. *Journal of Structural Engineering-ASCE,* **111**(4), 939-943.

Simpson, J. R. & Montgomery, D. C. 1998. A performance-based assessment of robust regression methods. *Communications in Statistics-Simulation and Computation,* **27**(4), 1031-1049.

Skanavis, V. 2013. *Geological and anthropogenic factors in the formation of coastline: the case study of Varkiza — S. Attica.* Technical Univeristy of Crete.

Sklar, A. 1959. Fonctions de répartition á n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris,* **8**, 229-231.

Smith, R. L. & Weissman, I. 1994. Estimating the extremal index. *Journal of the Royal Statistical Society Series B-Methodological,* **56**(3), 515-528.

Soares, C. G. & Henriques, A. C. 1996. Statistical uncertainty in long-term distributions of significant wave height. *Journal of Offshore Mechanics and Arctic Engineering-Transactions of the Asme,* **118**(4), 284-291.

Solari, S. & Losada, M. A. 2016. Simulation of non-stationary wind speed and direction time series. *Journal of Wind Engineering and Industrial Aerodynamics,* **149**, 48-58.

Song, M. X., Chen, K., Zhang, X. & Wang, J. 2016. Optimization of wind turbine micro-siting for reducing the sensitivity of power generation to wind direction. *Renewable Energy,* **85**, 57-65.

Soukissian, T. 2013. Use of multi-parameter distributions for offshore wind speed modeling: The Johnson S-B distribution. *Applied Energy,* **111**, 982-1000.

Soukissian, T., Papadopoulos, A., Skrimizeas, P., Karathanasi, F., Axaopoulos, P., Avgoustoglou, E., et al. 2017. Assessment of offshore wind power potential in the

Aegean and Ionian Seas based on high-resolution hindcast model results. *Aims Energy,* **5**(2), 268-289.

Soukissian, T. H. 2014. Probabilistic modeling of directional and linear characteristics of wind and sea states. *Ocean Engineering,* **91**, 91-110.

Soukissian, T. H. & Arapi, P. M. 2011. The effect of declustering in the r-largest maxima model for the estimation of Hs -design values. *The Open Ocean Engineering Journal,* **4**, 34-43.

Soukissian, T. H. & Chronis, G. T. 2000. POSEIDON: marine environmental, monitoring, forecasting and information system for the Greek Seas. *Mediterranean Marine Science,* **1**(1), 71-78.

Soukissian, T. H., Chronis, G. T. & Nittis, K. 1999. POSEIDON: Operational marine monitoring system for Greek seas. *Sea Technology,* **40**(7), 31-37.

Soukissian, T. H. & Kalantzi, G. D. 2009. A New Method for Applying r-Largest Maxima Model for Design Sea-State Prediction. *International Journal of Offshore and Polar Engineering,* **19**(3), 176-182.

Soukissian, T. H. & Karathanasi, F. E. 2017. On the selection of bivariate parametric models for wind data. *Applied Energy,* **188**, 280-304.

Soukissian, T. H., Karathanasi, F. E. & Voukouvalas, E. G. 2014. Effect of outliers in wind speed assessment. *The Twenty-fourth International Ocean and Polar Engineering Conference.* Busan, Korea: International Society of Offshore and Polar Engineers.

Soukissian, T. H. & Papadopoulos, A. 2015a. Effects of different wind data sources in offshore wind power assessment. *Renew Energ,* **77**, 101-114.

Soukissian, T. H. & Papadopoulos, A. 2015b. Effects of different wind data sources in offshore wind power assessment. *Renewable Energy,* **77**, 101-114.

Soukissian, T. H. & Tsalis, C. 2015. The effect of the generalized extreme value distribution parameter estimation methods in extreme wind speed prediction. *Natural Hazards,* **78**(3), 1777-1809.

Soulsby, R. 1997. *Dynamics of marine sands: A manual for practical applications*, Telford.

Soulsby, R. L. & Damgaard, J. S. 2005. Bedload sediment transport in coastal waters. *Coastal Engineering,* **52**(8), 673-689.

Srivastava, M. S. 1995. Comparison of the inverse and classical estimators in multi-univariate linear calibration. *Communications in Statistics-Theory and Methods,* **24**(11), 2753-2767.

Staudte, R. G. & Sheather, S. J. 1990. *Robust estimation and testing*, Wiley.

Stevens, E. W. 1992. Uncertainty of extreme flood estimates incorporating historical data. *Water Resources Bulletin,* **28**(6), 1057-1068.

Stewart, D. A. & Essenwanger, O. M. 1978. Frequency distribution of wind speed near the surface. *Journal of Applied Meteorology,* **17**(11), 1633-1642.

Stopa, J. E. & Cheung, K. F. 2014. Intercomparison of wind and wave data from the ECMWF Reanalysis Interim and the NCEP Climate Forecast System Reanalysis. *Ocean Modelling,* **75**, 65-83.

Stopa, J. E., Cheung, K. F., Tolman, H. L. & Chawla, A. 2013. Patterns and cycles in the Climate Forecast System Reanalysis wind and wave data. *Ocean Modelling,* **70**, 207-220.

Swart, D. H. 1974. *Offshore sediment transport and equilibrium beach profile,* Delft Hydraul. Lab. Publ. 131, Delft Univ. of Technol., Delft, Netherlands.

Tajvidi, N. 2003. Confidence intervals and accuracy estimation for heavy-tailed Generalized Pareto distributions. *Extremes,* **6**(2), 111-123.

Tang, H. & Lin, J. Z. 2013. Early-stage evolution of particle size distribution with Johnson's S-B function due to Brownian coagulation. *Physica Scripta,* **87**(5).

Tavner, P. 2017. *Wave and tidal generation devices: reliability and availability,* London, UK, The Institution of Engineering and Technology.

Taylor, C. C. 2008. Automatic bandwidth selection for circular density estimation. *Computational Statistics & Data Analysis,* **52**(7), 3493-3500.

Tellinghuisen, J. 2000. Inverse vs. classical calibration for small data sets. *Fresenius Journal of Analytical Chemistry,* **368**(6), 585-588.

Thennadil, S. N., Dewar, M., Herdsman, C., Nordon, A. & Becker, E. 2018. Automated weighted outlier detection technique for multivariate data. *Control Engineering Practice,* **70**, 40-49.

Todorovic, P. 1978. Stochastic Models of Floods. *Water Resources Research,* **14**(2), 345-356.

Tolman, H. L. 1997. User manual and system documentation of WAVEWATCH-III version 1.15. NOAA/NWS/NCEP OMB Technical Note Nr. 151.

Tous, M. & Romero, R. 2013. Meteorological environments associated with medicane development. *International Journal of Climatology,* **33**(1), 1-14.

Trucano, T. G., Swiler, L. P., Igusa, T., Oberkampf, W. L. & Pilch, M. 2006. Calibration, validation, and sensitivity analysis: What's what. *Reliability Engineering & System Safety,* **91**(10-11), 1331-1357.

Tukey, J. W. 1960. A survey of sampling from contaminated distributions. *In:* Olkin, I., Ghurye, S., Hoeffding, W., Madow, W. G. & Mann, H. B. (eds.) *Contributions to Probability and Statistics.* Stanford: Stanford University Press.

Tzavelas, G. 2009. Maximum likelihood parameter estimation in the three-parameter gamma distribution with the use of Mathematica. *Journal of Statistical Computation and Simulation,* **79**(12), 1457-1466.

Van Rijn, L. C. 1993. *Principles of sediment transport in rivers, estuaries and coastal seas,* Amsterdam, The Netherlands, Aqua Publications.

Vanem, E. 2016. Joint statistical models for significant wave height and wave period in a changing climate. *Marine Structures,* **49**, 180-205.

von Mises, R. 1918. Über die 'Ganzzahligkeit' der Atomgewichte und verwandte Fragen. *Physikalische Zeitschrift,* **19**, 490-500.

von Mises, R. 1936. La distribution de la plus grande de n valeurs. [Reprinted in *Selected Papers II*, American Mathematical Society, Providence R.I. (1954), 271-294].

Wakazuki, Y. 2013. Modified Relative Humidity Based on the Johnson's S-B Distribution Function. *Sola,* **9**, 111-114.

Wald, A. 1940. The fitting of straight lines if both variables are subject to error. *Annals of Mathematical Statistics,* **11**, 284-300.

Wang, F. K. & Keats, J. B. 1995. Improved percentile estimation for the two-parameter Weibull distribution. *Microelectronics Reliability,* **35**(6), 883-892.

Webb, N. P., McGowan, H. A., Phinn, S. R., Leys, J. F. & McTainsh, G. H. 2009. A model to predict land susceptibility to wind erosion in western Queensland, Australia. *Environmental Modelling & Software,* **24**(2), 214-227.

Wei, K., Arwade, S. R., Myers, A. T., Valamanesh, V. & Pang, W. C. 2017. Effect of wind and wave directionality on the structural performance of non-operational offshore wind turbines supported by jackets during hurricanes. *Wind Energy,* **20**(2), 289-303.

Wentworth, C. K. 1922. A scale of grade and class terms for clastic sediments. *Journal of Geology,* **30**(5), 377-392.

Wheeler, S. & Watson, G. S. 1964. A distribution-free two-sample test on circle. *Biometrika,* **51**(1-2), 256-257.

Wiberg, P. L. & Sherwood, C. R. 2008. Calculating wave-generated bottom orbital velocities from surface-wave parameters. *Comput Geosci-UK,* **34**(10), 1243-1262.

Williams, E. J. 1969. A Note on Regression Methods in Calibration. *Technometrics,* **11**(1), 189-192.

Williams, J. J. & Rose, C. P. 2001. Measured and predicted rates of sediment transport in storm conditions. *Marine Geology,* **179**(1), 121-133.

Willmott, C. J. 1982. Some comments on the evaluation of model performance. *Bulletin of the American Meteorological Society,* **63**(11), 1309-1313.

Wisnowski, J. W., Montgomery, D. C. & Simpson, J. R. 2001. A comparative analysis of multiple outlier detection procedures in the linear regression model. *Computational Statistics & Data Analysis,* **36**(3), 351-382.

Withers, C. S. & Nadarajah, S. 2015. Estimating trend from seasonal data: is daily, monthly or annual data best? *Environmetrics,* **26**(7), 488-501.

World Meteorological Organization 2017. *WMO Guidelines on the Calculation of Climate Normals*, (WMO-No. 1203). 2017 ed. Geneva.

Wu, X. N., Hu, Y., Li, Y., Yang, J., Duan, L., Wang, T. G., et al. 2019. Foundations of offshore wind turbines: A review. *Renewable & Sustainable Energy Reviews,* **104**, 379-393.

Yanalagaran, R. & Ramli, N. I. 2018. Assessment of coastal erosion related to wind characteristics in peninsular Malaysia. *Journal of Engineering Science and Technology,* **13**(11), 3677-3690.

Yohai, V. J. 1987. High breakdown-point and high-efficiency robust estimates for regression. *Annals of Statistics,* **15**(2), 642-656.

Young, I. R., Sanina, E. & Babanin, A. V. 2017. Calibration and cross validation of a global wind and wave database of altimeter, radiometer, and scatterometer measurements. *Journal of Atmospheric and Oceanic Technology,* **34**(6), 1285-1306.

Zaman, A., Rousseeuw, P. & Orhan, M. 2001. Econometric applications of high-breakdown robust regression techniques. *Economics Letters,* **71**(1), 1-8.

Zecchetto, S. & Cappa, C. 2001. The spatial structure of the Mediterranean Sea winds revealed by ERS-1 scatterometer. *International Journal of Remote Sensing,* **22**(1), 45-70.

Zecchetto, S. & De Biasio, F. 2007. Sea surface winds over the Mediterranean basin from satellite data (2000-04): Meso- and local-scale features on annual and seasonal time scales. *Journal of Applied Meteorology and Climatology,* **46**(6), 814-827.

Zhai, J. J., Yin, Q. L. & Dong, S. 2017. Metocean design parameter estimation for fixed platform based on copula functions. *Journal of Ocean University of China,* **16**(4), 635-648.

Zhang, H., Yu, Y. J. & Liu, Z. Y. 2014. Study on the Maximum Entropy Principle applied to the annual wind speed probability distribution: A case study for observations of intertidal zone anemometer towers of Rudong in East China Sea. *Applied Energy,* **114**, 931-938.

Zhang, H. M., Bates, J. J. & Reynolds, R. W. 2006. Assessment of composite global sampling: Sea surface wind speed. *Geophysical Research Letters,* **33**(17).

Zhang, J., Chowdhury, S., Messac, A. & Castillo, L. 2013. A multivariate and multimodal wind distribution model. *Renewable Energy,* **51**, 436-447.

Zhang, L. D., Li, Q. K., Guo, Y. J., Yang, Z. L. & Zhang, L. 2018. An investigation of wind direction and speed in a featured wind farm using joint probability distribution methods. *Sustainability,* **10**(12), 4338.

Zhang, L. J., Packard, K. C. & Liu, C. M. 2003. A comparison of estimation methods for fitting Weibull and Johnson's S-B distributions to mixed spruce-fir stands in northeastern North America. *Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere,* **33**(7), 1340-1347.

Zhou, J. Y., Erdem, E., Li, G. & Shi, J. 2010. Comprehensive evaluation of wind speed distribution models: A case study for North Dakota sites. *Energy Conversion and Management,* **51**(7), 1449-1458.

Zhou, S., Shan, B., Xiao, Y., Li, C., Hu, G., Song, X., et al. 2017. Directionality effects of aligned wind and wave loads on a Y-shape semi-submersible floating wind turbine under rated operational conditions. *Energies,* **10**, 2097.

Curriculum Vitae

Mrs Flora KARATHANASI (flora@mail.ntua.gr, fkarathanasi@hcmr.gr): Rural and Surveying engineer with MSc in Marine Technology and Science (NTUA). Her postgraduate thesis is entitled: "Comparative study of alternative regression methods and application to wind data". She is a PhD candidate of SNAME, NTUA, working on her thesis entitled "Probabilistic modelling of linear and directional wind and wave data with applications to the marine environment" since 2014, under the supervision of Prof. K. Belibassakis. She has considerable research experience, working at the Hellenic Centre for Marine Research (HCMR), serving fixed-term positions since 2014. She has participated in 8 European and National funded research projects mainly related with assessment studies of marine renewable energy resources (including offshore wind, waves and currents) in the Mediterranean Sea. She has co-authored 13 papers in journals and 9 contributions in scientific conferences. Scopus Citations: 136, h-index 7. GoogleScholar 130 citations, h-index 7, i10-index 4.

<u>Academic Degrees</u>: Diploma in Rural and Surveying Engineering from National Technical University of Athens (2009), Degree: 7.23/10.00. Master of Science in Marine Technology and Science from the School of Naval Architecture and Marine Engineering of the same University (2013), Degree: 8.53/10.00 (3rd out of 19 total postgraduate students). Master Thesis title: "Comparative study of alternative regression methods and application to wind data" (10/10).

<u>Current Position</u>: Ph.D. candidate, School of Naval Architecture & Marine Engineering, National Technical University of Athens. Dissertation title: "Probabilistic modelling of linear and directional wind and wave data with applications to the marine environment". Working for GLAFKI as a Research Assistant, Institute of Oceanography, Hellenic Centre for Marine Research.

<u>Publications</u>: Author and/or co-author of 13 international journal publications and 9 conference papers published in the corresponding proceedings.
Selected papers in scientific journals:

KARATHANASI, F., Belibassakis, K., 2019. A cost-effective method for estimating long-term effects of waves on beach erosion with application to Sitia bay, Crete. Oceanologia 61(2): 276–290.

Soukissian, T., KARATHANASI, F., Axaopoulos, P., Voukouvalas, E.G., V. Kotroni, V., 2018. Offshore wind climate analysis and variability in the Mediterranean Sea. International Journal of Climatology 38: 384–402.

Belibassakis, K., KARATHANASI, F., 2017. Modelling nearshore hydrodynamics and circulation under the impact of high waves at the coast of Varkiza in Saronic-Athens Gulf. Oceanologia 59(3): 350–36.

Soukissian, T.H., Denaxa, D., KARATHANASI, F., Prospathopoulos, A., Sarantakos, K., Iona A., Georgantas, K., Mavrakos, S., 2017. Marine Renewable Energy in the Mediterranean Sea: Status and Perspectives. Energies 10(10): 1512.

Soukissian, T., Papadopoulos, A., Skrimizeas P., KARATHANASI, F., Axaopoulos, P., Avgoustoglou, E., et al., 2017. Assessment of offshore wind power potential in the Aegean and Ionian Seas based on high-resolution hindcast model results. AIMS Energy 5(2): 268–289.

Soukissian, T.H., KARATHANASI, F.E., 2017. On the selection of bivariate parametric models for wind data. Applied Energy 188: 280–304.

Soukissian, T.H., KARATHANASI, F.E., 2016. On the use of robust regression methods in wind speed assessment. Renewable Energy 99: 1287–1298.

Soukissian, T., KARATHANASI, F., Axaopoulos, P., 2016. Satellite-based offshore wind resource assessment in the Mediterranean Sea. IEEE Journal of Oceanic Engineering, 42(1): 7387.

Soukissian, T., Reizopoulou, S., Drakopoulou, P., Axaopoulos, P., KARATHANASI, F., Fraschetti, S., et al., 2016. Greening offshore wind with the Smart Wind Chart evaluation tool. Web Ecology 16(1): 73–80.

Papers in conferences (review based on full paper or abstract):

KARATHANASI, F.E., Soukissian, T.H., Axaopoulos, P.G., 2016. Calibration of wind directions in the Mediterranean Sea, Proceedings of the 26th International Ocean and Polar Engineering Conference, 1: 491-497, Rhodes, Greece, June 26–July 1.

Soukissian, T., Reizopoulou, S., Drakopoulou, P., Axaopoulos, P., KARATHANASI, F., Bray, L., Kyriakidou, C., Papathanassiou, E., Papadopoulos, A., Voukouvalas, E., 2015. Offshore wind farms and Marine Protected Areas: integration in the context of the FP7 COCONET project, International Conference "Ecology at the Interface, Rome, September 21–25, organized by the European Ecological Federation.

Soukissian, T., KARATHANASI, F. Voukouvalas, E., 2014. Effect of outliers in wind speed assessment, Proceedings of the 24th International Offshore (Ocean) and Polar Engineering Conference, 1: 362-369, Busan, June 15–20.

Selected papers in conferences (review based on abstract):

Soukissian, T., KARATHANASI, F., Prospathopoulos, A., Mavrakos, S., Konispoliatis, D., 2018. Wind and wave power potential offshore Rhodes Isl., 5th International Interdisciplinary Symposium Poiiesa, Rhodes, Greece, October 5–7.

KARATHANASI, F., Belibassakis, K., Anagnostou, C., 2017. Simulation of wave field and sediment transport at the Sitia Bay, 7th National Congress on Upgrading and Management of Coastal Zones, pp. 33–42, Athens, Greece, November 20–22.

Anagnostou, C., Belibassakis, K., KARATHANASI, F., 2017. Coastal Erosion in the Sitia Crete Bay - Rehabilitation of the Coast Based on Nourishment Techniques as an Alternative to Hard Work Interventions, 7th National Congress on Upgrading and Management of Coastal Zones, pp. 421–430, Athens, Greece, November 20–22.

Soukissian, T., Axaopoulos, P., KARATHANASI, F., 2015. Offshore wind power in the Mediterranean Sea using Blended Sea Winds, OCEANS'15 International Conference, MTS/IEEE, Genova, May 18–21, ISBN: 978-1-4673-7164-3.

Reviewer of Oceanologia (Elsevier), Ocean Engineering (Elsevier), Journal of Marine Science and Application (Springer), International Journal of Energy Research (Wiley), International Society of Offshore and Polar Engineers (ISOPE), SNAME Maritime Convention.

Scholarship from the Special Account for Research Grants, National Technical University of Athens (02/2017–02/2018).