Prodotto n. T2.2.1

Progetto - Projet

SPlasH! - Stop alle Plastiche in H2O!

PRODOTTO T2.2.1: Report sulla costruzione degli scenari climatologici e di rischio

LIVRABLE T2.2.1: Rapport sur la construction de scénarios climatologiques et de risques

Partner responsabile - Partner responsable: Università di Genova

Partner contributori - Partenaires contributeurs: Université de Toulon, European Research Institute

Nome del prodotto	Redatto da:	Verificato da:	Validato da:
T2.2.1 - Report sulla costruzione degli scenari climatologici e di rischio	G. Cremonini	G. Besio	A. Stocchino
Data:	20/05/2020	25/05/2020	28/06/2020

Descrizione del Prodotto: Report metodologico per l'individuazione di scenari di forzanti ambientali per la simulazione dei processi di dispersione delle micro-plastiche in ambito portuale e nelle acque circostanti.

Description du livrable : Rapport méthodologique pour l'identification de scénarios de forçages environnementaux pour la simulation des processus de dispersion des microplastiques dans les eaux portuaires et les eaux environnantes.

Sintesi

La realizzazione di simulazioni numeriche della dinamica della dispersione delle microplastiche nelle acque portuali e nelle aree circostanti i porti considerati nell'area di progetto, richiede l'identificazione di specifiche condizioni meteo-marine (forzanti). La definizione di scenari multi-variabile risulta di per sé di grande difficoltà in quanto le forzanti considerate risultano essere numerose ed estremamente variabili nel tempo. In particolare, siccome l'interesse è rivolto alla dinamica della circolazione marittima, è necessario definire delle condizioni specifiche per quanto riguarda l'intensità e la direzione del vento, l'altezza, il periodo e la direzione del moto ondoso, il valore della pressione atmosferica, l'oscillazione e l'eventuale presenza di rivi e fiumi negli specchi d'acqua oggetto dello studio. L'identificazione di scenari specifici viene quindi realizzata tramite tecniche di data mining basate su diversi algoritmi di clusterizzazione noti in letteratura. L'analisi viene eseguita su un set di diverse variabili e considerando una finestra temporale caratteristica per il fenomeno preso in esame. Tale finestra temporale può differire significativamente dai tempi caratteristici delle singole fenomenologie studiate. È possibile identificare un numero limitato di scenari rappresentativi delle condizioni meteo-marine tipiche di un dato paraggio, ottimizzando in questa maniera il carico computazionale delle simulazioni numeriche.

Synthèse

La réalisation de simulations numériques de la dynamique de la dispersion des microplastiques dans les eaux portuaires et dans les eaux environnantes des ports appartenant à la zone du

projet nécessite l'identification de conditions météo-marines spécifiques (forçages). La définition de scénarios à variables multiples est difficile à réaliser compte tenu du nombre de facteurs de forçage et de leur variabilité dans le temps. Plus précisément, dans la mesure où l'objet des simulations est la dynamique de la circulation des eaux marines, il est nécessaire de définir des conditions spécifiques relatives à l'intensité et la direction du vent, la hauteur, la période et la direction de la houle, la valeur de la pression atmosphérique, l'oscillation et l'éventuelle présence de cours d'eau et fleuves affluant dans les masses d'eau qui font l'objet de notre étude. L'identification de scénarios spécifiques est par conséquent réalisée à travers des techniques d'exploration de données (*data mining*) fondées sur différents algorithmes de partitionnement (*clustering*). L'analyse est effectuée sur un ensemble de différentes variables et sur une durée pertinente pour le phénomène examiné. Cette durée peut varier considérablement par rapport aux durées caractéristiques de chaque phénomène singulier. Il est possible d'identifier un nombre limité de scénarios représentatifs des conditions météo-marines caractéristiques d'une zone déterminée, et d'optimiser ainsi la charge computationnelle des simulations numériques.

Indice

1 Introduzione

L'analisi dei processi fisici relativi alla circolazione costiera e ai processi di dispersione in tale zona è sempre stata di grande interesse da un punto di vista sia scientifico sia applicativo. La necessità di avere a disposizione degli strumenti di analisi e predizione di tale tipo di processi è sentita non solo in ambito di pianificazione e programmazione delle attività umane nella fascia costiera (i.e. controllo della dispersione dei dragaggi, progettazione degli emissari degli impianti di depurazione) ma anche in ambito di gestione delle emergenze e della qualità delle acque (i.e. sversamenti e dispersione di inquinanti sia lato terra che lato mare, incidenti in mare ed operazioni di search&rescue). In quest'ottica nelle ultime decadi, soprattutto grazie alla crescita esponenziale della potenza di calcolo dei moderni computer, è stato fatto sempre maggiore affidamento all'utilizzo di programmi di calcolo in grado di realizzare simulazioni numeriche dell'idrodinamica marina e costiera e dei processi di dispersione sia da un punto di vista euleriano (dinamica della concentrazione) che da un punto di vista lagrangiano (dispersione di massa e oggetti). In particolare, è stata sviluppata una vasta gamma di modelli numerici per l'analisi, la simulazione e la risoluzione di problemi di fluidodinamica: si tratta di modelli di Fluidodinamica Computazionale (CFD) che permettono di simulare i fenomeni ambientali più importanti risolvendo numericamente le leggi fisiche dei fluidi. Il principale utilizzo della CFD, infatti, è quello di risolvere equazioni per via numerica in situazioni reali e complesse.

In questo modo è possibile ad esempio analizzare processi di morfodinamica, studiando l'evoluzione costiera in seguito a mareggiate intense e gli eventuali fenomeni di erosione. La CFD permette di studiare la circolazione marina in modo tale da approfondire la conoscenza delle correnti che caratterizzano una determinata zona e affrontare potenziali problemi legati alla dispersione di inquinanti. Esistono inoltre modelli che permettono anche la valutazione del comportamento di strutture costiere, in fase di progettazione, soggette a particolari condizioni di mare.

Lo studio di tali fenomeni generalmente si basa su un'enorme quantità di informazioni che richiedono elevati tempi di computazione e grandi potenze di calcolo non sempre disponibili: questo, in particolare, può succedere quando ad esempio il dataset in esame proviene da un servizio di reanalisi climatologica, caratterizzato da alta risoluzione temporale e spaziale. In questo caso, può essere conveniente ridurre il numero di condizioni ambientali da tenere in considerazione per le simulazioni numeriche in modo da individuare e conservare i modi più significativi della variabilità del fenomeno. Risolvere un numero limitato di condizioni ambientali, altresì dette "scenari", è vantaggioso perché non solo permette di selezionare gli scenari più importanti per il processo investigato ma anche perché viene ridotto significativamente il carico computazionale necessario per risolvere l'intera catena di modellazione.

A tale fine, è possibile impiegare delle tecniche di "Data mining" ovvero di analisi massiva dei dati a disposizione, tramite algoritmi di clusterizzazione ("clustering"); tale approccio si è rivelato molto utile: infatti, permette di raggruppare un insieme di dati in classi di oggetti (*cluster*) sulla base della loro similarità/dissimilarità. Un cluster rappresenta un raggruppamento di elementi che sono simili tra loro e sono dissimili dagli elementi di un altro cluster. Il risultato che si ottiene è un sottoinsieme di elementi in grado di riassumere il dataset iniziale, mantenendo le sue proprietà principali.

Nella letteratura scientifica vi sono diverse applicazioni delle tecniche di clustering per l'identificazione di condizioni ambientali con differenti specifici obiettivi che riguardano non solo l'analisi di processi fisici ma anche la loro simulazione numerica. L'applicazione di tali algoritmi a database di reanalisi delle condizioni meteomarine (solitamente estremamente ampi per quanto riguarda la numerosità delle variabili archiviate, soprattutto nel caso di alta risoluzione spaziale e temporale) consente di descrivere le condizioni meteo-marine, selezionando determinati stati rappresentativi della sua variabilità, con l'obiettivo di implementarli in una metodologia di propagazione del moto ondoso (Camus, Mendez, Medina, & Cofiño, 2011b). Per la definizione degli stati di mare, gli autori considerano istanti orari delle serie temporali di altezza d'onda, periodo e direzione media, senza tener conto della loro evoluzione nel tempo.

Diverso è l'approccio di (Bárcena, Camus, García, & Álvarez, 2015) che utilizza le tecniche di clustering nell'ambito della simulazione dell'idrodinamica tridimensionale degli estuari ad alta risoluzione spaziale e temporale: viene infatti definita inizialmente una finestra temporale in base alla quale analizzare i dati di partenza, al fine di ottenere dei cluster rappresentati da definite serie temporali. L'impiego di tale approccio permette quindi di rappresentare sinteticamente la variabilità delle maree astronomiche e a individuare scenari delle forzanti in gioco per ottenere il comportamento medio riducendo la dimensione del dataset iniziale.

Emerge che la scelta della lunghezza della finestra temporale di dati da esaminare dipende non solo dal tempo scala delle forzanti considerate ma anche dal tempo caratteristico del processo che si vuole esaminare in seguito. Ad esempio, la propagazione del moto ondoso dal largo verso riva viene studiata considerando stati di mare orari, mentre il tempo scala tipico di problemi di dispersione è dell'ordine delle settimane.

Altra particolarità delle tecniche descritte nei lavori summenzionati è il modo in cui viene applicata la clusterizzazione: da una parte (Camus, Mendez, Medina, & Cofiño, 2011b) clusterizza le variabili considerandole in maniera congiunta, dall'altra (Bárcena, Camus, García, & Álvarez, 2015) effettua una clusterizzazione per ogni variabile coinvolta nell'analisi in maniera indipendente.

Tuttavia, lo studio di processi in ambito meteomarino non permette di considerare le variabili indipendentemente l'una dall'altra: in questo lavoro, infatti, proponiamo una metodologia che permetta di caratterizzare il clima marino considerando non solo le caratteristiche del moto ondoso ma anche la velocità del vento, il campo di pressione e la forzante di marea. Per la costruzione del dataset iniziale viene quindi scelta una finestra temporale adeguata in funzione del tipo di processo che si vuole studiare, ovvero la descrizione della dispersione di inquinanti/sedimenti/particelle in acque costiere, in seguito all'immissione in mare di una portata definita in un certo intervallo di tempo.

2 Dati & Metodi

2.1 Parametri Mete-Ocean

Le variabili meteo-marine impiegate nel presente studio derivano dai prodotti di hindcast del Dipartimento di Ingegneria Civile, Chimica ed Ambientale dell'Università di Genova (DICCA), www3.dicca.unige.it/meteocean/hindcast.html. Tramite una re-analisi delle condizioni atmosferiche, è stato costruito un database contenente dati orari di onda, vento e campo barico definiti su una griglia con risoluzione approssimativamente di 10 km lon/lat, estesa a tutto il bacino del Mar Mediterraneo (Mentaschi, Besio, Cassola, & Mazzino, Developing and validating a forecast/hindcast system for the Mediterranean Sea., 2013; Mentaschi, Besio, Cassola, & Mazzino, Performance evaluation of wavewatch iii in the mediterranean sea., 2015). L'implementazione del dataset di hindcast è avvenuta in seguito alla validazione e alla ottimizzazione della catena di modelli numerici impiegata (WRF per la parte meteo e WaveWatchIII per la parte onde) e ad oggi questi dati sono stati utilizzati in numerose ricerche e applicazioni (Re, Manno, Ciraolo, & Besio, 2019), (Leo, Besio, Zolezzi, & Bezzi, 2019), (Sartini, Besio, Dentale, & Reale, 2016), (De Girolamo, et al., 2018), (Sartini, Besio, & Cassola, 2017), (Zughayar, Gudmestad, De Leo, & Besio, 2017), (Mucerino, et al., 2019), (Besio, Briganti, Romano, Mentaschi, & Girolamo, 2017), (Ferretti, et al., 2018). Per lo sviluppo degli algoritmi di identificazione di scenari climatici caratteristici vengono prese in considerazione le serie temporali dal 1979 al 2018, su base oraria, di altezza d'onda significativa (H_s), di periodo e direzione di picco (T_p e θ_p, rispettivamente) delle componenti di velocità longitudinale/latitudinale del vento (w_x/w_y) e della pressione media sul livello del mare ($mslp$) in un punto griglia di fronte a Genova (Mar Tirreno, vedere Fig. 1). In un secondo momento, vengono ricavate le forzanti di marea (di seguito $\Delta\eta$) nella posizione selezionata grazie al Software di Previsione di Marea (TPXO.3) fornito dall'Università dello Stato dell'Oregon (Egbert & Erofeeva, 2002). L'escursione di marea è stata calcolata nello stesso intervallo di tempo e con la stessa frequenza per cui erano disponibili i dati meteomarini dell'hindcast.

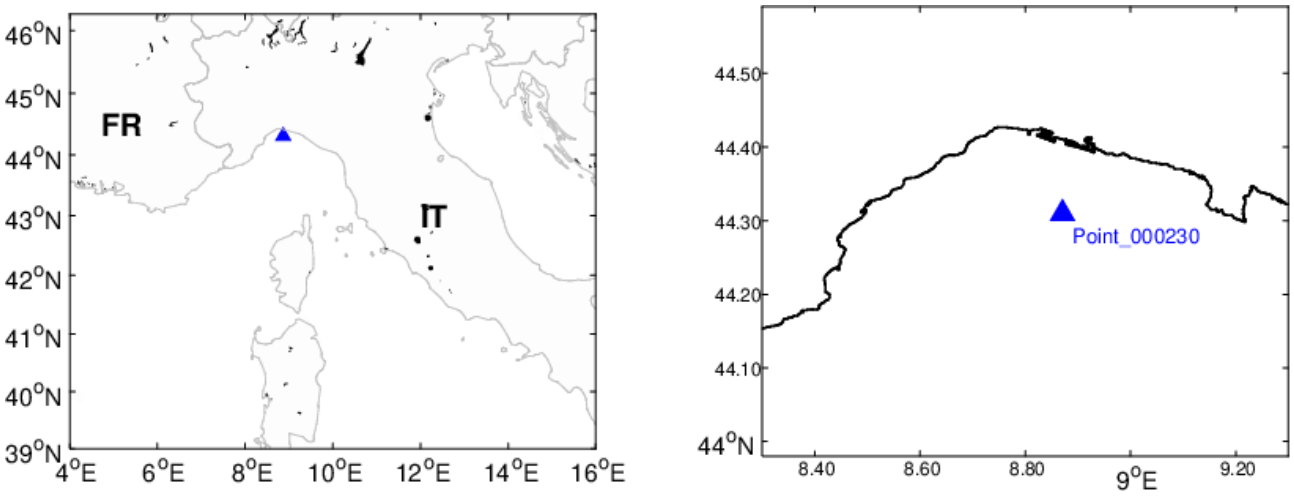

Figura 1: Zoom dell'area di studio. Il punto selezionato è evidenziato con un triangolo blu insieme al codice numerico del punto hindcast (lon/lat-8.8707/44.31; sistema di riferimento: WGS 84).

2.2 Analisi Cluster

La clusterizzazione è un insieme di tecniche di analisi multivariata dei dati volta alla selezione e al raggruppamento di elementi omogenei in un insieme di dati, basandosi su misure relative alla somiglianza tra gli elementi stessi, in termini di distanza in uno spazio multidimensionale. Può essere utilizzato per esaminare le distribuzioni dei dati, per osservare le caratteristiche di ciascuna distribuzione e per focalizzarsi su quelle di maggiore interesse. Alternativamente, esso può essere utilizzato come un passo di pre-processing dei dati per altri algoritmi che operano sui cluster individuati (come in questo caso).

Il punto di partenza delle tecniche di cluster di seguito spiegate è la modifica del dataset originale per la costruzione di una matrice di dati $X_{n,V}$, dove n e V sono rispettivamente il numero dei dati da modellare e delle variabili del problema. Dati V vettori di n dati, $X_{n,V}$ viene definita come:

$$X_{n,V} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,V} \\ \vdots & \ddots & \vdots \\ x_{1,n} & \cdots & x_{n,V} \end{bmatrix} \quad (1)$$

Il raggruppamento dei dati (o "clusterizzazione") a partire dal dataset così definito permette di selezionare un determinato numero di sottoinsiemi significativi, costituiti da righe della matrice X, indipendentemente dalla struttura temporale delle variabili stesse. Tale approccio permette di descrivere in maniera esaustiva la varianza delle forzanti attraverso un numero limitato di stati: infatti utilizzare l'intero dataset potrebbe talvolta diventare poco conveniente non solo da un punto di vista computazionale (nel caso di un numero troppo elevato di

simulazioni di condizioni meteo-marine) ma anche da un punto di vista di caratterizzazione di uno scenario tipo [si veda (Camus, Mendez F. J., & Medina, 2011a) per alcune applicazioni].

In questo lavoro, si fa riferimento alle tecniche di clustering partizionale che permettono di classificare i dati non essendo nota l'etichetta delle classi a priori. Vengono presi in considerazione il cosiddetto Maximum Dissimilarity Algorithm (d'ora in poi MDA) e il k-Means, due dei più diffusi algoritmi nel campo della Data Mining Analysis: permettono di partizionare i dati sulla base della loro dissimilarità/similarità e consentono di scegliere quali sono le caratteristiche di interesse per distinguere i vari gruppi (sulla base della scelta del numero di cluster).

2.2.1 MDA

L'obiettivo di MDA è di selezionare un sottoinsieme della matrice $X_{n,V}$, definito $X^*_{m,V}$ (matrice delle etichette, dove $m < n$), che meglio rappresenta la varianza complessiva dei dati. Il passaggio iniziale prevede la normalizzazione di tutte le variabili scalari (quindi, di ogni colonna della matrice $X_{n,V}$) in uno spazio comune, in modo tale da poter lavorare facilmente con variabili caratterizzate da diversi ordini di grandezza. L'algoritmo di selezione inizia con l'individuazione del primo dato significativo, identificato come lo stato più lontano dalla nuvola di punti iniziale, ovvero quello più "dissimilare" dal resto del dataset. Una volta selezionato, si procede con il calcolo della dissimilarità tra i vettori della matrice $X_{n-1,V}$ e il sottoinsieme $X^*_{1,V}$, quindi con il calcolo della distanza Euclidea così espressa:

$$d_{i1} = \|x_i - x_1^*\|, \;\; i = 1, \dots, n-1 \tag{2}$$

dove x e x^* indicano rispettivamente le righe delle matrici di input e di output. Il nuovo elemento viene successivamente selezionato come quello caratterizzato dal valore massimo di d_{i1} e aggiunto alla matrice target X^*. Come illustrato in (Camus, Mendez, Medina, & Cofiño, 2011b), si applica la versione MaxMin dell'algoritmo: infatti, in corrispondenza della k^{esima} iterazione ($k < m$, dove k è il numero di elementi della matrice X^*), la distanza da considerare per ogni elemento di $X_{n-k,V}$ è la minima rispetto a tutti i k vettori di $X^*_{k,V}$; tra tutte queste distanze viene selezionata quella massima e il corrispondente stato è di conseguenza aggiunto a X^*. Il calcolo finisce quando k equivale a m, ovvero una volta che viene raggiunto il numero di cluster (precedentemente stabilito). Alla fine del procedimento di costruzione della matrice X^*, si ottiene quindi un sottoinsieme di vettori che permette di riassumere il dataset di partenza con un numero minore di stati: gli elementi rimasti vengono assegnati al corrispondente vettore modello più vicino, formando delle vere e proprie classi di stati.

2.2.2 k-Means

Il k-Means è un algoritmo di clustering partizionale che permette di suddividere un insieme di oggetti (o vettori) in k gruppi, sulla base dei loro attributi. È una tecnica basata sul calcolo della distanza Euclidea tra i differenti elementi del dataset, come per il caso di MDA. Tuttavia, in questo caso l'obiettivo è quello di minimizzare la varianza intra-cluster, ognuno identificato con un centroide o punto medio della stessa dimensione dei dati originali (MacQueen, 1967). Partendo nuovamente da una normalizzazione delle variabili, l'algoritmo segue una procedura iterativa assegnando, al primo step, i centroidi in modo casuale (ovvero $X_{m,V}^{*,1}$), selezionati tra le righe della matrice $X_{n,V}$. Quindi, ogni dato (per esempio, la i[esima] riga di $X_{i,V}$, $i \in [1,n]$) viene "assegnato" al centroide più vicino:

$$m_i = m/\left(d_i = \|x_i - x_{m,V}^{*,1}\|, i = 1, \dots n - m\right) \quad (3)$$

dove m_i è l'i[esimo] cluster a cui appartengono i dati. Una volta che gli m gruppi sono stati definiti, vengono calcolati i nuovi centroidi ($x_{m,V}^{*,2}$) semplicemente come media dei rispettivi cluster:

$$x_{m,V}^{*,2} = \sum_{x_i \in m_j} \frac{x_i}{n_j} \quad (4)$$

essendo n_j il numero degli elementi appartenenti al j[esimo] cluster. Il procedimento di classificazione finisce quando la posizione dei centroidi non si modifica in maniera rilevante tra due successive iterazioni: in tal caso si dice che l'algoritmo è giunto a convergenza.

2.2.3 Selezione del numero ottimale di cluster

Le tecniche di clusterizzazione sopra descritte richiedono come primo passo fondamentale: definire un numero di clusters m appropriato. La scelta può essere fatta in modo soggettivo, se l'utente, ad esempio, desidera che i dati vengano assegnati a un determinato numero di classi, oppure se conosce come e con quali distribuzioni si presentano i dati (come nell'esempio riportato nella Sezione 2.5). Tuttavia, se il numero ottimale di cluster non è noto a priori, è necessario introdurre un'analisi di sensitività sui risultati della clusterizzazione. Ad esempio, (Bárcena, Camus, García, & Álvarez, 2015) fa riferimento all'indice CE (nell'ambito del Model Efficiency) proposto da (Nash & Sutcliffe, 1970), che permette di valutare l'efficienza degli stati del modello nella riproduzione del dataset di partenza. Approccio simile viene applicato in (Núñez, et al., 2019) dove viene calcolato il *Mean Skill Index* introdotto da (Willmott, 1981): tale coefficiente riflette la precisione con cui le variabili classificate si approssimano alle variabili originali.

In questo lavoro, si è preso in considerazione quanto suggerito da (Solari, et al., 2017), ovvero l'uso della "varianza totale", calcolata come:

$$W^2 = \frac{1}{n}\sum_{i=1}^{m}\sum_{i=1}^{m_j}\left(x_j - x_i^*\right)^2 \quad (5)$$

essendo n il numero totale degli stati e m_j il numero di elementi appartenenti al j^{esimo} cluster; x_i^* è l'i^{esimo} centroide della matrice X^*. L'equazione (5) può essere perciò usata, nel caso del k-Means, come un indicatore utile per valutare il numero di cluster, portando a un valore quasi-asintotico di W^2.

Un approccio simile può essere applicato nello schema di MDA, anche se in questo caso l'obiettivo di clusterizzazione è quello di spiegare complessivamente la varianza dei dati. Pertanto, in questo caso si fa riferimento al calcolo della distanza media tra gli stati modello, definita come:

$$\bar{d} = \frac{1}{m-1}\sum_{i=1}^{m-1} d_i \quad (6)$$

dove d_i è la distanza tra lo stato target i^{esimo} e (i+1)^{esimo} selezionati tramite il procedimento iterativo di MDA. Viene, quindi, definito m come quel valore oltre il quale non si apprezzino cambiamenti significativi di $\bar{d}$ o di W^2.

2.2.4 Clusterizzazione di dati in funzione dell'evoluzione temporale

Fino ad ora, il dataset iniziale considerato per l'applicazione degli algoritmi di clustering (per esempio, $X_{n,V}$), era definito semplicemente affiancando le serie storiche originali delle variabili, ciascuna caratterizzata da risoluzione temporale oraria. Ogni variabile, infatti, è rappresentata da un vettore e ogni suo elemento descrive un singolo stato.

Tuttavia, l'obiettivo di questo lavoro è quello di selezionare stati che devono essere impiegati per realizzare simulazioni numeriche di particolari fenomeni fisici: diventa importante, quindi, tenere in considerazione la variabilità e l'andamento temporale delle forzanti in gioco. Devono perciò essere selezionati scenari significativi dal punto di vista temporale, e ciò richiede di riordinare i dati in input mantenendo la struttura temporale delle forzanti in esame. In accordo con (Bárcena, Camus, García, & Álvarez, 2015), una volta definita la scala temporale delle forzanti, le rispettive serie di dati possono essere organizzate di conseguenza in serie di assegnata estensione, in modo tale da preservare l'andamento nel tempo delle variabili. L'obiettivo di tale riorganizzazione dei dati è proprio quello di cercare di mantenere l'impronta degli eventi effettivamente osservati. Detto nt il numero di passi temporali (cioè il numero di ore) di cui è composta la scala temporale di riferimento, il vettore x viene ridefinito nel modo seguente:

$$x'_{j,:} = x[i\delta + 1 : nt + i\delta], i = 0, \dots, (n - nt)/\delta, j = i + 1 \quad (7)$$

dove δ rappresenta lo shift tra due successivi intervalli di tempo (ovvero il numero di passi temporali tra i punti iniziali di due successive x').

Successivamente l'analisi di clusterizzazione può essere eseguita direttamente sulla matrice riorganizzata X'; in questo caso le righe della matrice non fanno più riferimento a stati puntuali delle diverse variabili, ma sono al contrario delle finestre temporali per ogni singola grandezza in esame. Si riporta nell'equazione 8 un esempio di tale operazione eseguito sulla serie temporale di una variabile X.

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{nt} \\ \vdots \\ x_n \end{bmatrix} \qquad X' = \begin{bmatrix} x_1 & \cdots & x_{nt} \\ x_{i\delta+1} & \cdots & x_{nt+i\delta} \\ \vdots & \cdots & \vdots \\ x_{n-nt+1} & \cdots & x_n \end{bmatrix} \tag{8}$$

2.3 Scelta della scala di tempo di riferimento

Come mostrato nella Sezione 2.2, la variabilità temporale delle forzanti esaminate gioca un ruolo fondamentale per la definizione della non stazionarietà degli scenari tipo. Si deve, infatti, selezionare un intervallo di tempo rappresentativo, che risulti idoneo per tutte le grandezze tenute in considerazione. Come confermato anche da (Bárcena, Camus, García, & Álvarez, 2015) la lunghezza della serie a breve termine che determina lo scenario tipo è legata alle scale temporali che caratterizzano il segnale. In particolare, gli autori stabiliscono una lunghezza scala per le forzanti di marea in funzione dei cicli di marea tipici dell'area geografica in esame; mentre utilizzano un indice di stima della durata degli impulsi di flusso per la valutazione del tempo scala della portata fluviale.

Per la nostra applicazione, si è deciso di valutare la scala temporale dei singoli parametri calcolando le loro rispettive funzioni di autocorrelazione (ACF): tale funzione permette di definire il grado di dipendenza tra i valori assunti da una variabile campionata nel suo dominio in ascissa. In altre parole, ACF rappresenta la correlazione incrociata tra il segnale all'istante t e un altro istante posto ad una certa distanza (*lag*), e permette dunque di verificarne la mutua dipendenza. L'obiettivo infatti è quello di determinare la frequenza fondamentale del segnale in esame. Dato un set di dati x di lunghezza *n,* la funzione di autocorrelazione per un determinato lag *k* è definita come:

$$\begin{cases} ACF_k = \frac{c_k}{c_0} \\ c_k = \frac{1}{n}\sum_{t=1}^{n-k}(x_t - \bar{x})(x_{t+k} - \bar{x}) \end{cases} \tag{9}$$

essendo c_k la covarianza per il lag *k*, c_0 la varianza campionaria delle serie storiche, mentre $\bar{x}$ è il valore medio della serie. La ACF di una particolare variabile permette di valutare il tempo scala su

cui, mediamente, una variabile decorre su stati significativi. La ACF viene valutata all'interno di un limite di confidenza (o l'intervallo di confidenza "IC") per un determinato livello di significatività α:

$$\begin{cases} IC_k = ACF_k \pm z_{\frac{\alpha}{2}} \times SE(ACF_k) \\ SE(ACF_k) = \sqrt{\frac{1+2\sum_{j=1}^{k-1} ACF_j^2}{n}} \end{cases} \quad (10)$$

dove $z_{\alpha/2}$ è il quantile relativo all'intervallo $[\alpha/2, 1-\alpha/2]$ nello spazio Normale, $SE(ACF_k)$ rappresenta l'errore standard stimato.

Per ogni grandezza, si valuta la funzione di autocorrelazione per una finestra lunga dieci giorni, composta da 240 lag: ogni lag è legato alla risoluzione temporale del dataset, che in questo caso è pari a un'ora. La finestra ha attraversato l'intero periodo considerato (1979-2018), con il ritardo iniziale spostato di un'ora alla volta. Infine, i valori risultanti della ACF sono stati mediati per ogni lag. α è stato scelto pari al 5%. I risultati dell'analisi appena descritta sono illustrati nella Fig. 2.

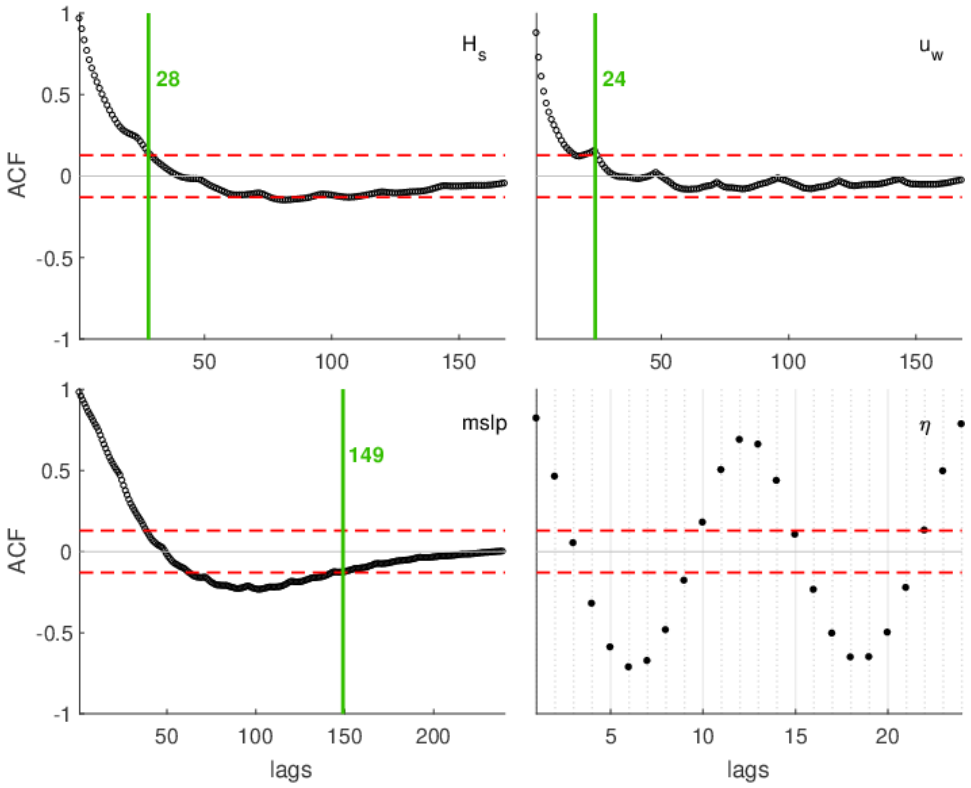

Figura 2: Funzione di Auto-Correlazione per alcune variabili meteo-marine impiegate nell'analisi.

Come mostra la Fig.2, la scala temporale per gli stati di mare equivale approssimativamente a un giorno sia per l'altezza d'onda H_s sia per la velocità del vento (facendo riferimento a u_w). Questo risultato riflette il clima marino locale del punto in esame (di fronte a Genova), essendo prevalentemente caratterizzato da wind waves (le cosiddette onde di mare vivo): esse, di solito, nascono, si sviluppano e svaniscono su scala giornaliera. Un'analoga considerazione può essere

applicata al periodo d'onda T_p, il quale è strettamente legato a H_s. D'altra parte, la funzione di autocorrelazione per *mslp* richiede quasi una settimana per raggiungere valori tra i rispettivi limiti dell'intervallo di confidenza (IC): molto probabilmente ciò è dovuto alle scale sinottiche che caratterizzano le traiettorie dei minimi di pressione. Infine, come ci si può aspettare, l'oscillazione della superficie libera è condizionata da maree semidiurne e maree di quarto turno, caratterizzate da cicli di circa 12 e 6 ore rispettivamente.

I risultati della ACF suggeriscono che le variabili coinvolte nell'analisi si caratterizzano su scale differenti, perciò è necessario stabilire uno schema comune in modo tale da facilitare l'applicazione delle analisi di clustering successive. In tali situazioni, la finestra modello di tempo (chiamata Δt^*) dovrebbe essere scelta in funzione del tipo di processo che deve essere simulato in un secondo momento. In ogni caso, Δt^* non dovrebbe essere minore della scala temporale più piccola relativa alle variabili in esame, affinché tutte possano essere adeguatamente caratterizzate sulle proprie scale rappresentative. Sulla base di tali considerazioni e tenendo presente i tempi scala tipici dei processi di dispersione, nel presente studio si è scelto di impostare Δt^* pari a una settimana.

2.4 Analisi di correlazione tra le variabili

Una volta definita la scala di tempo di riferimento, risulta necessario valutare la correlazione che caratterizza le variabili in esame. Come riportato nell'articolo di (Bárcena, Camus, García, & Álvarez, 2015), se le variabili non risultano mutuamente correlate è possibile applicare la clusterizzazione indipendentemente per ogni grandezza. D'altro canto, se la correlazione delle grandezze considerate non è trascurabile, queste devono essere considerate in modo congiunto.

Perciò, si prosegue con la valutazione delle correlazioni tra tutte le variabili prese in esame; per quanto riguarda le variabili circolari, (come ad esempio θ_p e la direzione del vento incidente θ_w) si introduce *l'indice di correlazione circolare* proposto da (Fisher & Lee, 1983):

$$r = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} sinsin\,(\theta_i-\theta_j)\,sin(\alpha_i-\alpha_j)}{\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(\theta_i-\theta_j)\,sin^2(\alpha_i-\alpha_j)}} \tag{11}$$

Per l'applicazione dell'equazione (10) sono stati presi in esame solo alcuni anni della serie storiche considerate: in particolare sono stati considerati solo tre anni (come esempio) caratterizzati da differenti intensità d'onda, in modo tale da avere un'idea del diverso andamento di r. Per ogni anno analizzato, è stata fissata un'altezza d'onda di soglia (H_{th}) e sono stati conservati solo la direzione di picco dell'onda θ_p e la direzione del vento θ_w legati alle altezze per cui si verifica la seguente condizione $H_s \geq H_{th}$. I risultati ottenuti dalle analisi di correlazione sono mostrati in Fig. 3 e Fig. 4.

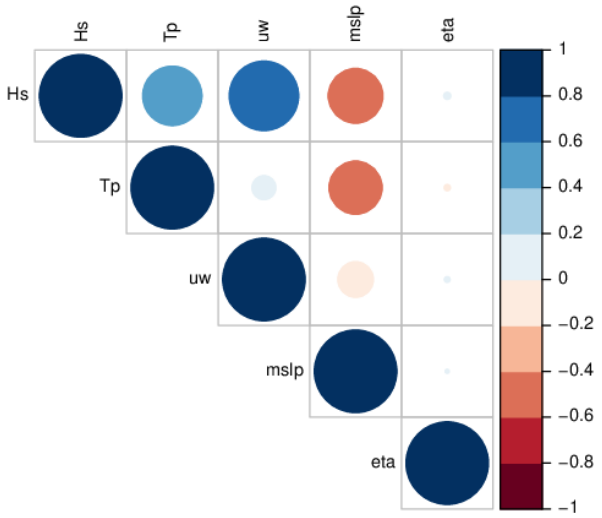

Figura 3: Correlazione delle grandezze non direzionali.

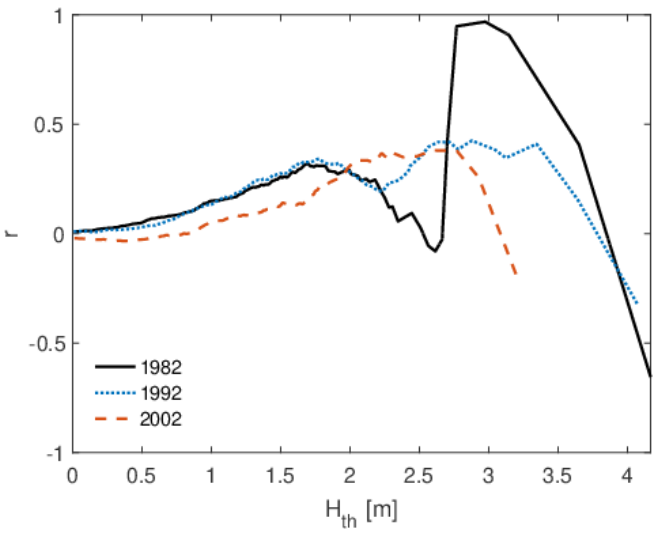

Figura 4: Correlazione circolare tra θ_p e θ_w per differenti anni e altezze d'onda di soglia.

Dall'analisi della Fig. 3 si può notare come H_s, T_p e u_w siano altamente correlate, mentre risultano anti-correlate nei confronti della *mslp*. Infatti, i sistemi di bassa pressione sono associati a intensi stati di mare, che sono a loro volta legati a elevate velocità del vento. Questo può essere dedotto inoltre dai valori di correlazione circolare tra le direzioni d'onda e di vento; come mostrato dalla Fig. 4, r raggiunge valori significativi in corrispondenza di condizioni di mareggiata, mentre per stati di mare estremi si annulla fino a diventare negativo, indicando anti-correlazione tra le direzioni di mare e di vento incidente. Ciò è verosimilmente dovuto al fatto che gli stati del mare estremamente intensi non sono guidati dal vento, ma sono legati a onde lunghe, generate lontano dalla zona di campionamento, e che viaggiano per lunghe distanze (le cosiddette "swell").

Infine, per quanto riguarda $\Delta\eta$, non vi è alcuna prova di correlazione significativa con le altre variabili considerate, poiché i cicli di marea astronomica non influenzano il clima delle onde né

dipendono dalla pressione media sul livello del mare. Pertanto, le oscillazioni della superficie libera potrebbero essere raggruppate indipendentemente dagli altri parametri che, al contrario, devono essere elaborati nel loro insieme.

2.5 Clusterizzazione di variabili circolari

Le tecniche di clusterizzazione dei dati, spiegate nella Sez. 2.2 comportano il calcolo delle distanze Euclidee tra gli elementi di un dataset e alcuni elementi target definiti nello stesso spazio, con l'obiettivo di raggruppare i dati in accordo con il relativo algoritmo impiegato.

Tuttavia, nel momento in cui vengono utilizzate variabili circolari (come la direzione di propagazione delle onde), bisogna prendere in considerazione una precauzione che permetta di gestire tali dati facilmente e di evitare errori in presenza di discontinuità nello spazio delle variabili. Per spiegare meglio tale concetto, si può fare riferimento a θ_p, definita in accordo con la convenzione nautica (le direzioni sono definite in senso orario, partendo da Nord). Ad esempio, si possono considerare due onde provenienti da Nord caratterizzate da direzioni di arrivo rispettivamente di 0° e 360°: in tale caso, la prima è più simile (ovvero più vicina) ad un'onda target che si propaga verso ovest, mentre la seconda è più vicina ad un'onda orientata verso est (essendo i rispettivi θ_p uguali a 90° e a 270° rispettivamente). Ciononostante, le due condizioni d'onda considerate presentano la stessa direzione di provenienza e per questo motivo un raggruppamento differente sarebbe insensato. Per risolvere questo problema, viene di solito applicata una correzione direttamente alle direzioni. Definite θ_1 e θ_2 le direzioni delle due onde in esame:

$$\Delta\theta = \begin{cases} 2\pi - (\theta_1 - \theta_2), & (\theta_1 - \theta_2) > \pi \\ (\theta_1 - \theta_2), & (\theta_1 - \theta_2) \in [-\pi \div \pi] \\ (\theta_1 - \theta_2) + 2\pi, & (\theta_1 - \theta_2) < -\pi \end{cases} \tag{12}$$

Tuttavia, in questo lavoro viene introdotto ed impiegato un approccio più adeguato. Letteralmente, facendo ancora ricorso alle direzioni di arrivo delle onde di esempio, θ_p può essere proiettata lungo l'asse est/nord come segue:

$$\begin{cases} \theta_{p,x} = -\cos(\theta_p - 90) \\ \theta_{p,y} = sin(\theta_p - 90) \end{cases} \tag{13}$$

La correzione applicata agli argomenti delle funzioni sinusoidali è necessaria per far in modo che le componenti proiettate siano coerenti con la convenzione nautica: ad esempio, le componenti sono entrambe positive (negative) nel terzo (primo) quadrante, mentre sono di segni discordanti nel secondo e quarto quadrante. Dall'Eq. 13, si deduce che $\theta_{p,x}$ e $\theta_{p,y}$ possono raggiungere valori negativi, i quali ovviamente sono privi di senso dal punto di vista fisico, ma significativi ai fini della

clusterizzazione, in quanto consentono di raggruppare i dati considerando sia le informazioni sull'intensità delle onde che le direzioni di provenienza.

La figura 5 mostra un sintetico esempio di un clima d'onda bimodale, costituito da due insiemi di dati d'onda con direzioni di provenienza distribuite intorno a 0° e 180°, rispettivamente. È stato applicato l'algoritmo del k-Means prima mantenendo H_s e θ_p, e poi utilizzando le proiezioni nel piano cartesiano. I risultati sono messi a confronto nella Fig. 6.

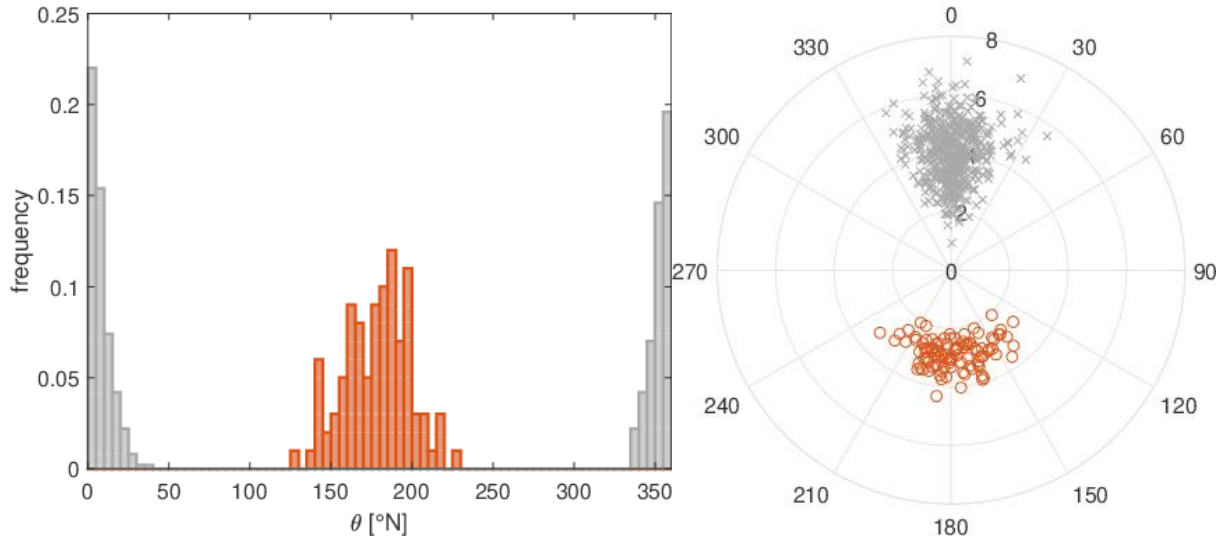

Figura 5. Esempio di un clima d'onda bimodale: distribuzione delle frequenze (sinistra) e grafico polare (destra).

Quando il raggruppamento degli stati viene eseguito prendendo in considerazione θ_p e H_s separatamente, e senza correzione sulle distanze circolari, il raggruppamento dei dati viene influenzato dalla direzione di provenienza delle onde, con conseguente classificazione errata. In questo caso, le onde sono divise tra quelle appartenenti ai quadranti 1-2 (croci grigie) e ai quadranti 3-4 (cerchi arancioni, riquadro sinistro di Fig. 6). Al contrario, se il clustering viene applicato a $\theta_{p,x}$ e $\theta_{p,y}$, la classificazione ha esito positivo: il carattere bimodale del clima ondoso viene correttamente rilevato (pannello di destra di Fig. 6).

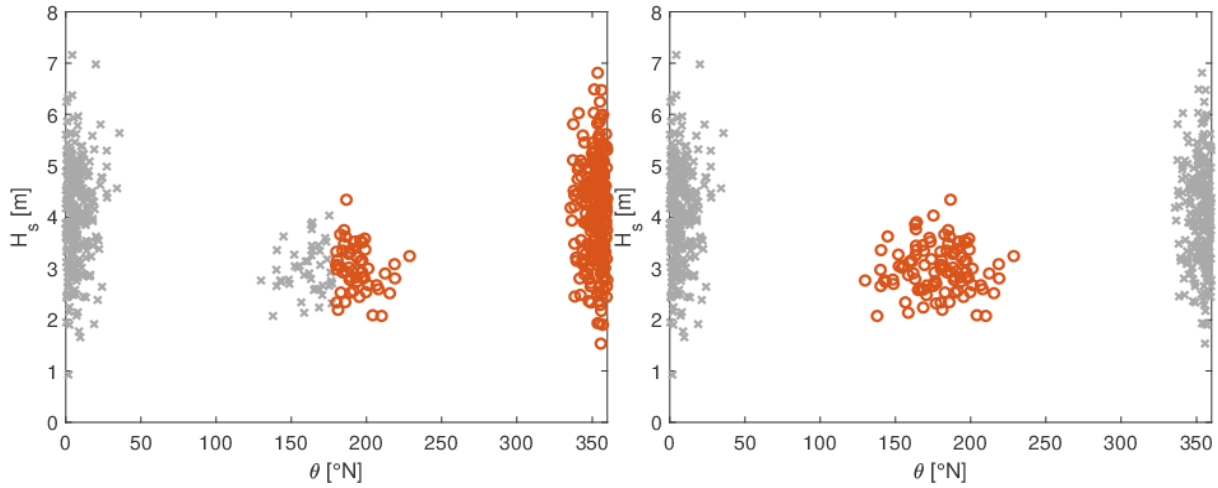

Figura 6. Clustering di moto ondoso bimodale: variabili originali (destra), proiezioni (sinistra).

2.6 Clusterizzazione regionale

Talvolta le simulazioni numeriche che vengono condotte in ambito marittimo non sono limitate all'analisi di un singolo punto, ma sono rivolte allo studio del comportamento di un'area geografica più estesa (ad esempio, un'area davanti a un porto o un golfo). Si propone quindi un ulteriore sviluppo della metodologia qui illustrata: è la clusterizzazione regionale che permette di tenere in conto della variabilità spaziale dei parametri meteomarini. L'obiettivo è quello di ottenere un'unica classificazione delle condizioni meteomarine di una certa area geografica considerando le informazioni estese a un intero sottobacino. In questo modo è possibile riassumere accuratamente il clima marino del sottobacino mantenendo poche decine di stati che sono in grado di esprimere la variabilità climatica della zona in esame.

La clusterizzazione regionale prevede quindi di applicare gli algoritmi k-means e MDA ad un dataset che contiene le stesse grandezze dell'analisi precedente ($H_s, \theta_{p,x}, \theta_{p,y}, T_p, w_x, w_y, mslp, \Delta\eta$) di tutti i punti hindcast appartenenti al sottobacino considerato. Si faccia riferimento alla mappa in Fig. 6 per visualizzare la localizzazione dei punti hindcast analizzati.

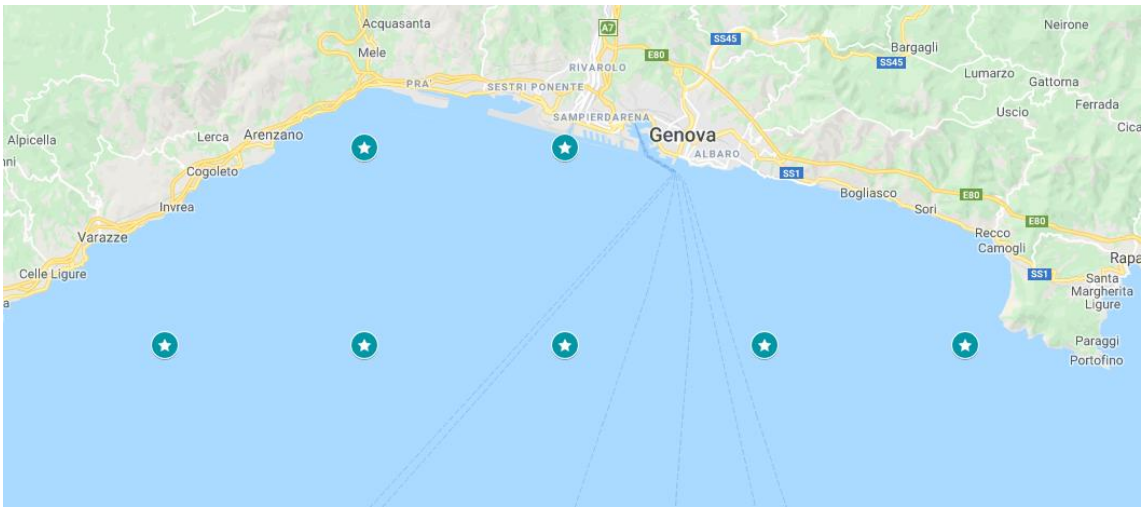

Figura 6: Zoom del sottobacino di interesse. I 7 punti considerati sono segnati con stelline blu.

3 Risultati

In accordo con l'analisi della correlazione e della scala temporale delle variabili, la matrice dei dati viene costruita considerando la lunghezza della finestra temporale nt pari a 168 elementi (= 7*24 ovvero, pari al numero delle ore in una settimana) per ogni variabile in esame: quindi vengono concatenate le matrici x' di $H_s, \theta_{p,x}, \theta_{p,y}, T_p, w_x, w_y, mslp, \Delta\eta$ (vedere Eq.7). Si ottiene, quindi, una matrice di dimensioni pari a $V \times nt$, dove V è il numero delle variabili analizzate (pari a 8).

La matrice X' è stata poi standardizzata lungo ogni colonna, centrando i dati intorno allo zero (sottraendo la media) e scalandoli rispetto alla deviazione standard:

$$X'_{:,j} = \frac{X'_{:,j} - \mu(X'_{:,j})}{\sigma(X'_{:,j})} \tag{14}$$

dove μ e σ rappresentano rispettivamente la media e la deviazione standard della j^{esima} colonna della matrice X'. In realtà, tale operazione permette all'algoritmo applicato successivamente di trattare sia dati negativi sia positivi, caratterizzati da differenti ordini di grandezza, evitando che le variabili più rilevanti alterino i successivi calcoli.

La Figura 8 mostra i risultati dell'analisi di sensitività eseguita sui dati, per l'algoritmo sia di MDA sia di k-Means (Equazione 5 e 6 rispettivamente). I valori complessivi di W^2 e $\bar{d}$ sono stati scalati nel range 0-1, dal momento che la scala originale non assume alcun rilievo, essendo le statistiche calcolate a partire da variabili standardizzate; la forma della curva è la sola caratteristica importante. In entrambi i casi è difficile scegliere un esatto numero di cluster, tale che il tasso di variazione per $\bar{d}$ e W^2 sia trascurabile; ciononostante, risulta evidente una pendenza più mite a partire dai 20/30 cluster. Come tale, per i seguenti calcoli m viene scelto pari a 30 per entrambi gli

algoritmi considerati, in modo da essere in grado di confrontare correttamente i risultati delle differenti tecniche.

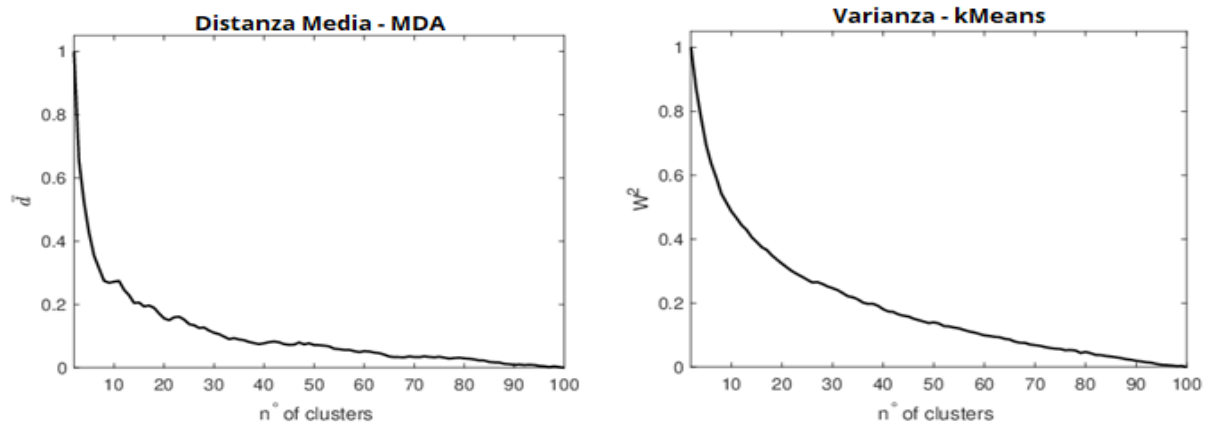

Figura 8. Analisi di sensitività in funzione del numero di cluster; algoritmo di MDA (pannello di sinistra) e algoritmo di k-means (pannello di destra).

In seguito, nelle Figure 9-12 sono riportati due esempi (scelti tra tutti i trenta cluster) di stati significativi ottenuti attraverso l'applicazione sia di k-Means sia di MDA. Nelle figure 13-16 vengono riportati come esempio due dei trenta cluster individuati dalla clusterizzazione areale. In particolare, sono stati scelti gli stati rappresentativi di un evento di libeccio e di tramontana risultanti da entrambe le analisi, in modo tale da avere un confronto migliore.

Vale la pena notare che le direzioni θ_p degli stati target sono stati ottenuti riproiettando le variabili $\theta_{p,x}$ e $\theta_{p,y}$, risultanti dalle analisi di clusterizzazione.

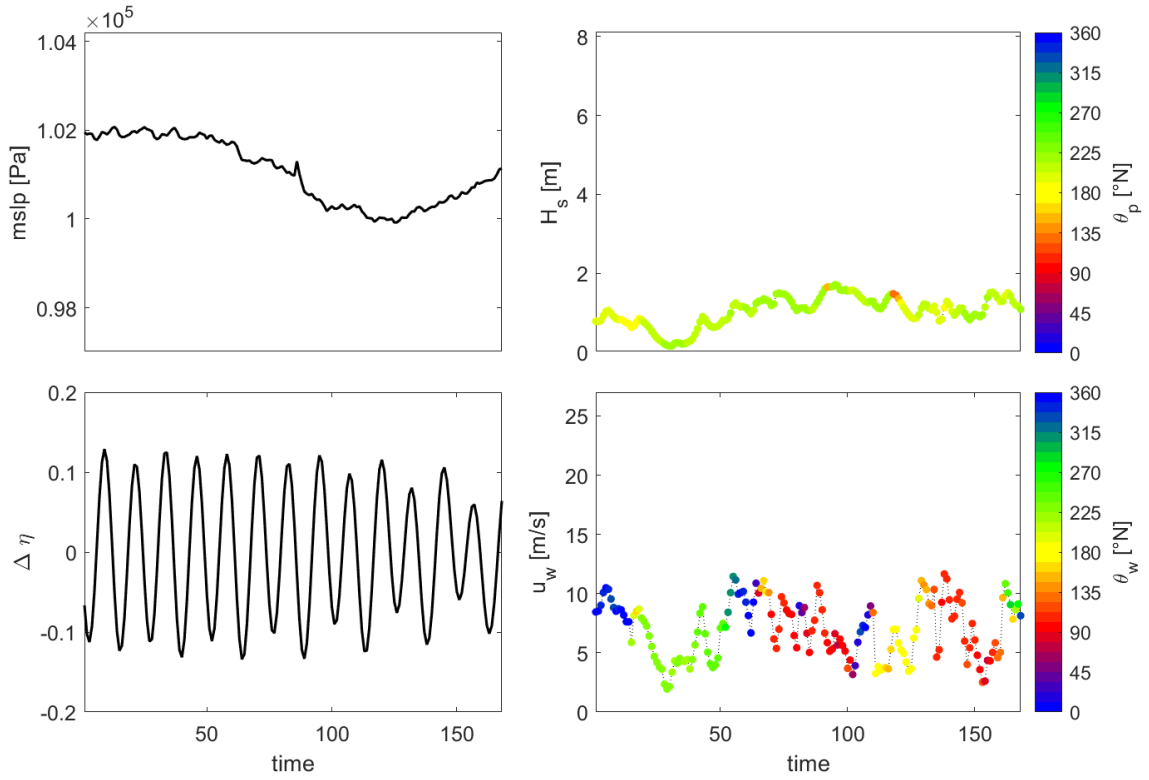

Figura 9. Stato significativo: caratteristiche di un evento di libeccio ottenuto tramite l'analisi k-means.

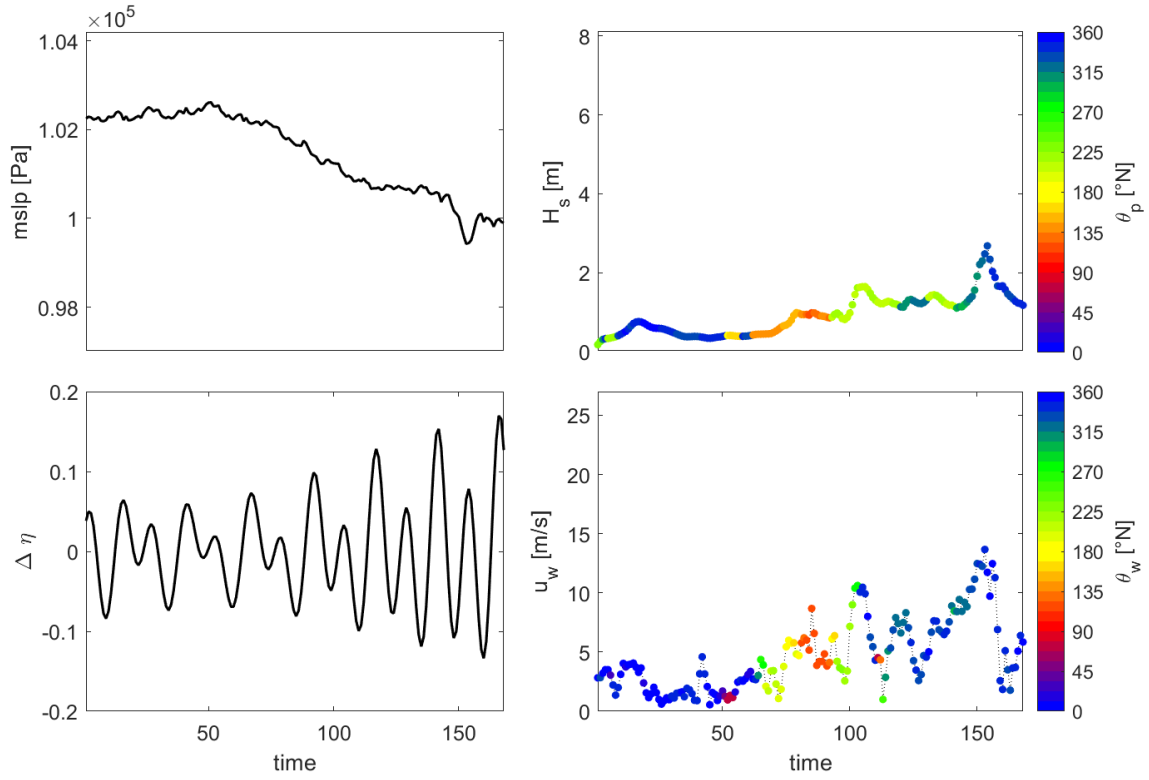

Figura 10. Stato significativo: caratteristiche di un evento misto di tramontana ottenuto tramite l'analisi k-means.

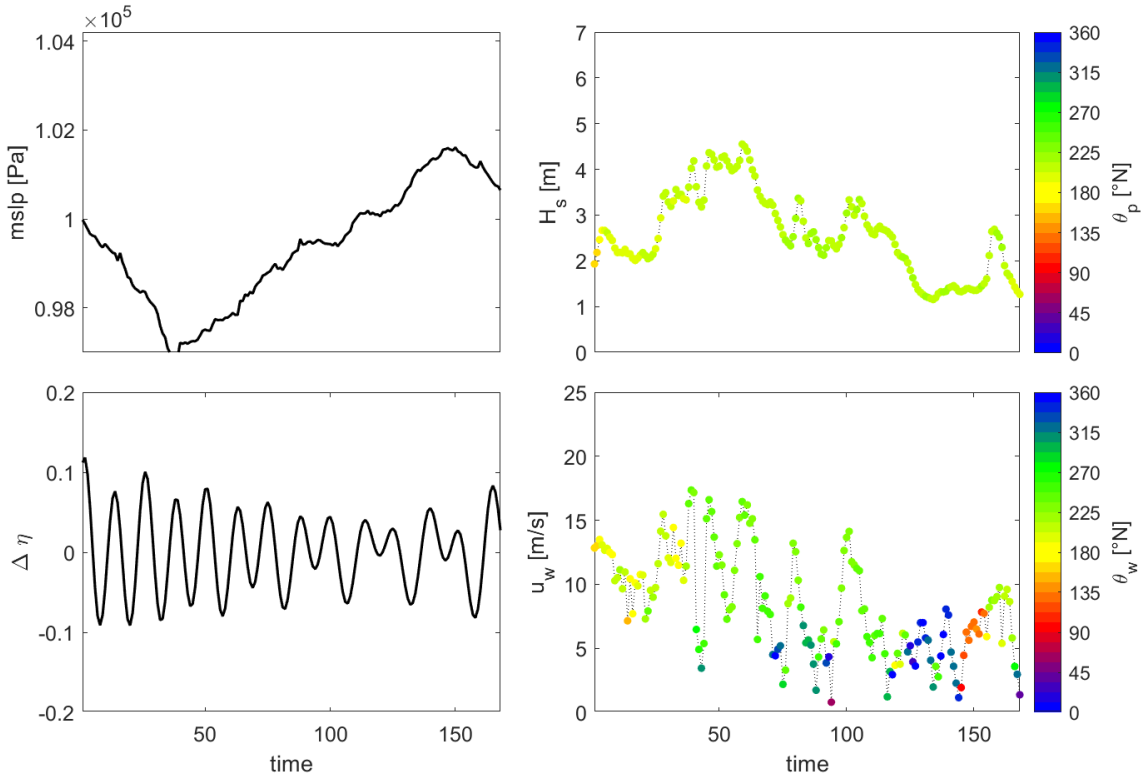

Figura 11. Stato significativo: caratteristiche di un evento di libeccio ottenuto tramite l'analisi MDA.

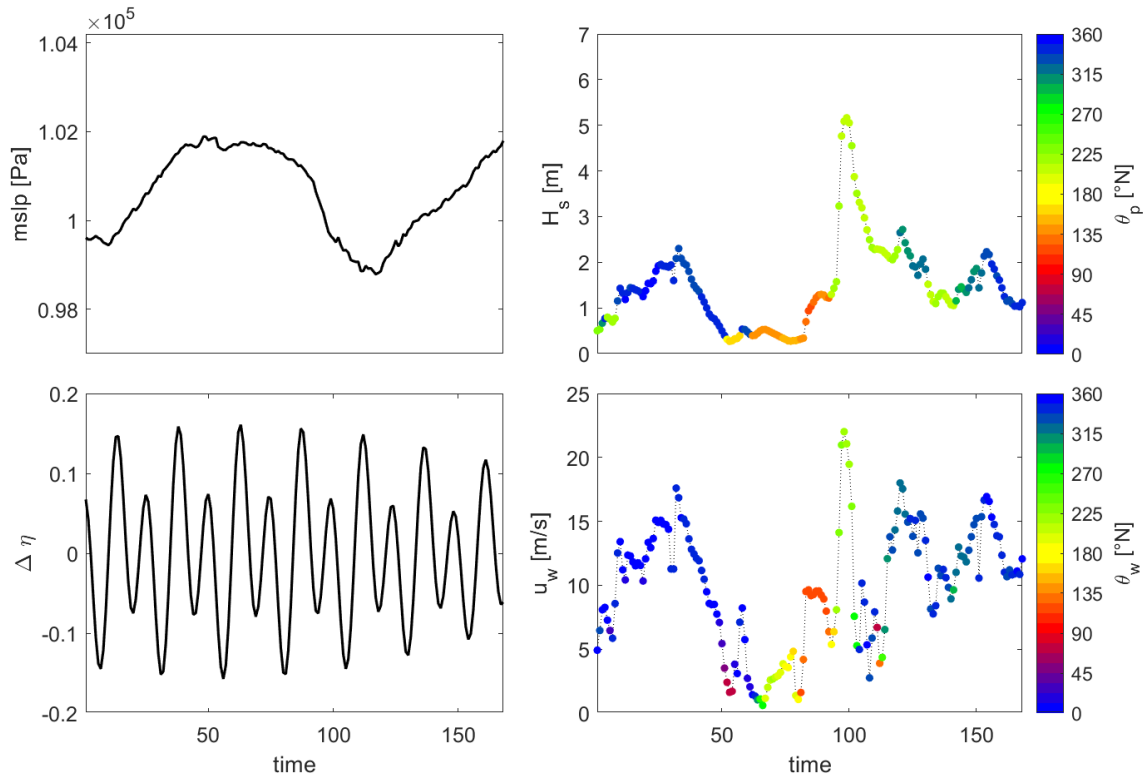

Figura 12. Stato significativo: caratteristiche di un evento misto di tramontana ottenuto tramite l'analisi MDA.

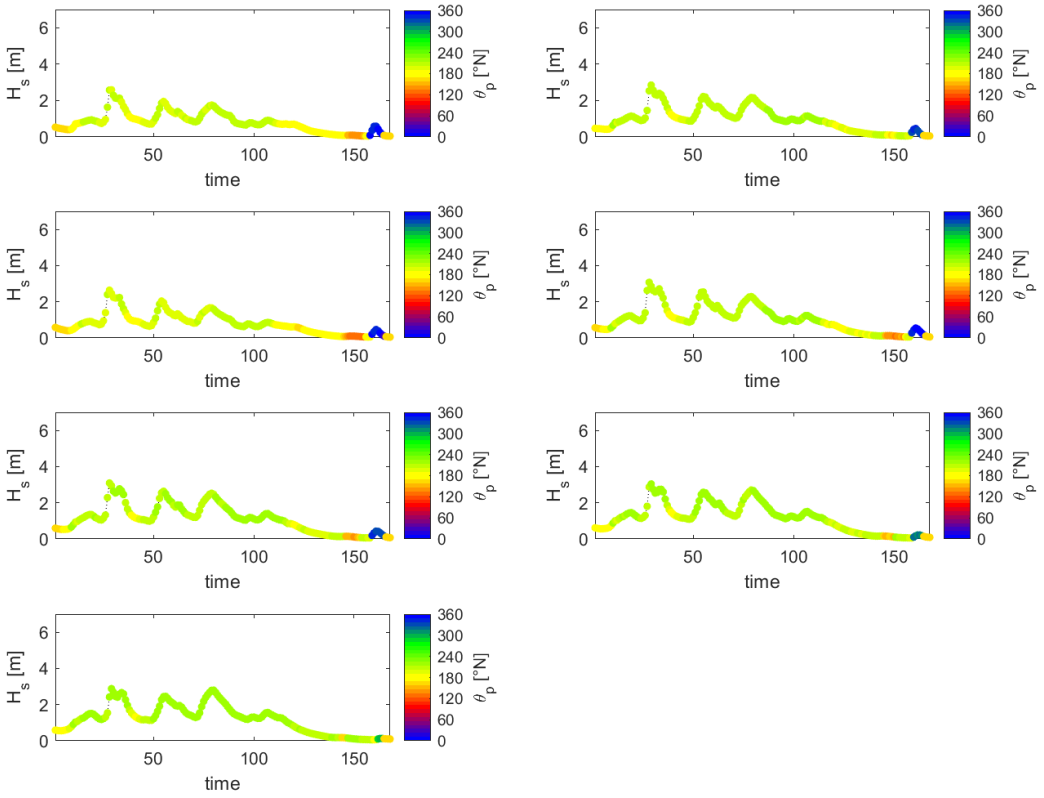

Figura 13. Stato significativo: caratteristiche di un evento di libeccio ottenuto tramite k-means regionale.

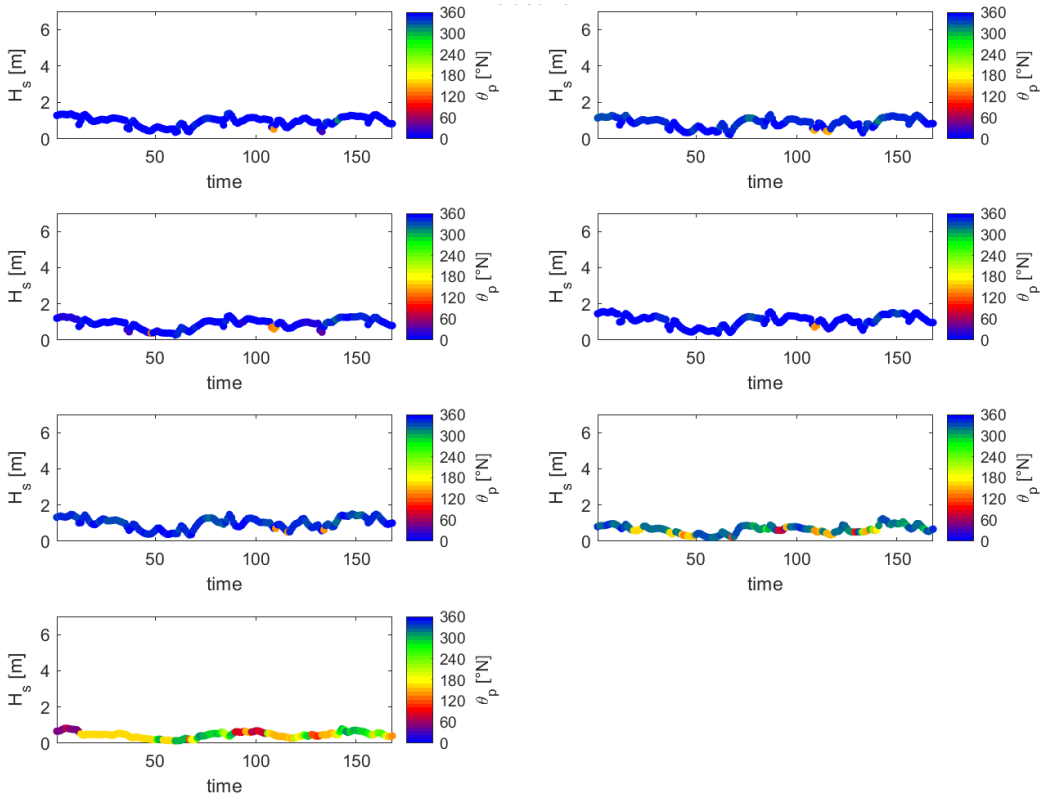

Figura 14. Stato significativo: caratteristiche di un evento di tramontana ottenuto tramite k-means regionale.

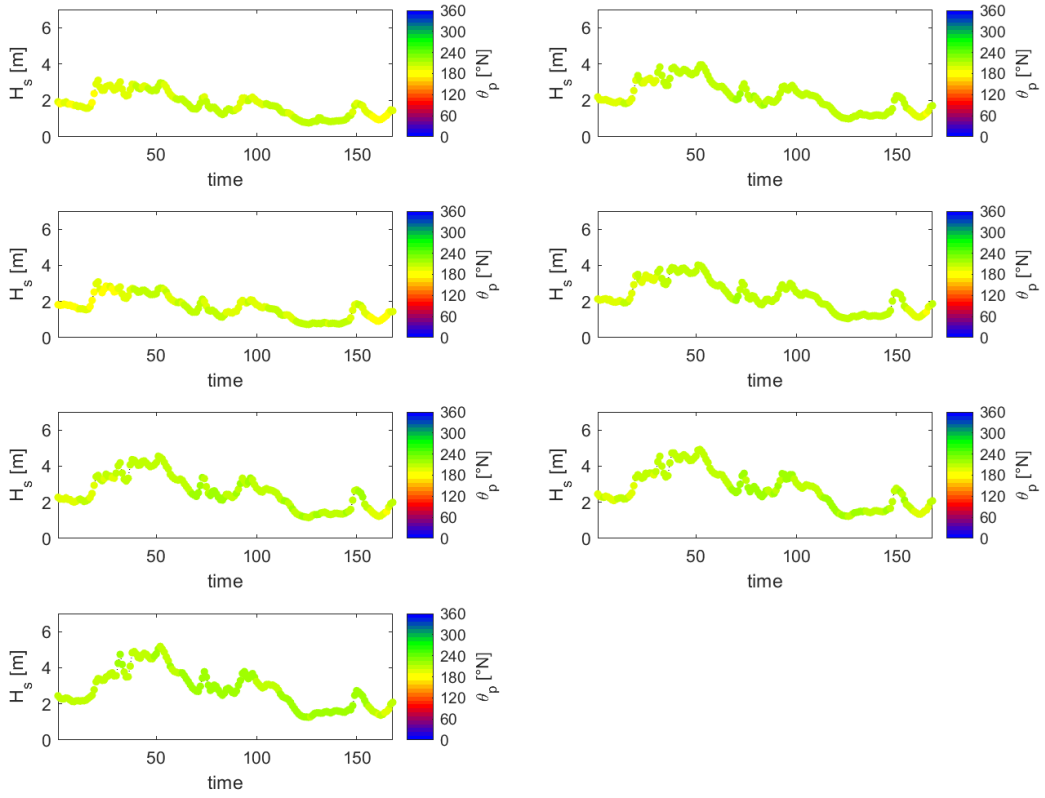

Figura 15. Stato significativo: caratteristiche di un evento di libeccio ottenuto tramite MDA regionale.

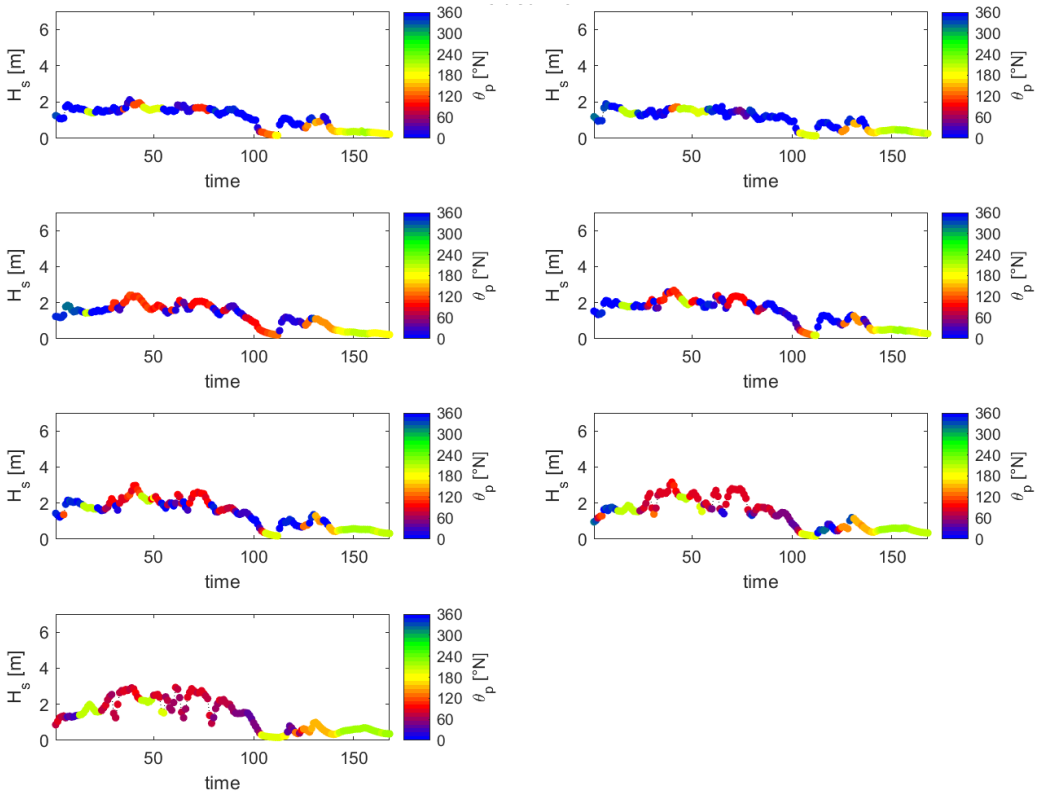

Figura 16. Stato significativo: caratteristiche di un evento misto di tramontana ottenuto tramite MDA regionale.

4 Discussione

Gli stati modello risultanti dall'analisi clustering dei dati catturano efficacemente le principali modalità del clima ondoso locale. Le figure 9-12 mostrano un paio di esempi di eventi di libeccio e di tramontana (ovvero onde che si propagano da SW e da N), che sono i due settori prevalenti del clima locale dell'area investigata in questo lavoro. Tale affermazione risulta evidente osservando la rosa direzionale di Fig. 17, che riporta la distribuzione direzionale di H_s in funzione di differenti classi di intensità.

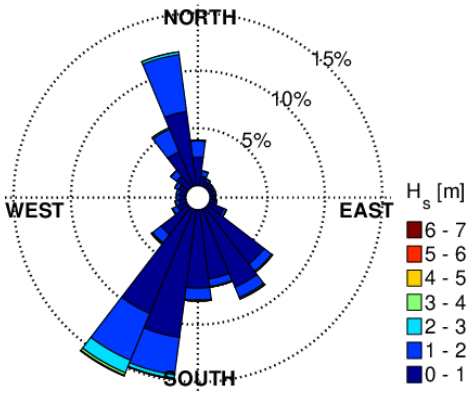

Figura 17. Grafico polare di Hs per il Point_000230 del database hindcast (vedere Fig. 1).

Ciò si riflette di conseguenza sulle caratteristiche degli stati modello: infatti la maggior parte degli stati realizzati sia con k-means sia con MDA mostra caratteristiche coerenti con questi settori, sia in termini di direzione di provenienza delle onde che di direzione del vento. Le figure 10 e 12 suggeriscono un'altra considerazione: osservando le serie temporali delle diverse variabili, si può notare come l'intensità dell'altezza d'onda sia proporzionale alla velocità del vento; al contrario, una significativa anti-correlazione caratterizza i profili di $H_s - u_w$ e quelli della $mslp$. Ciò non sorprende, poiché correlazioni simili tra l'intera serie delle variabili erano già state evidenziate nelle Fig. 3 e 4. Si precisa che le serie modello di T_p non sono riportate per chiarezza e semplicità, ma risultano fortemente proporzionali a quelle di H_s.

Infine, non è stata trovata alcuna correlazione apprezzabile tra le serie di dati di altezza d'onda, velocità del vento e pressione per l'evento di libeccio ottenute attraverso l'analisi del k-means (Fig. 9). Ciò è probabilmente dovuto al fatto che l'intensità dello scenario di mareggiata individuato non è così rilevante, anche se talvolta le mareggiate di libeccio di fronte a Genova sono caratterizzate da altezze d'onda fino a 6-7 m. Lo stato selezionato dall'algoritmo non si riferisce a una condizione estrema e burrascosa e, quindi, non è determinato da velocità del vento elevate né da sistemi di bassa pressione.

Da questa considerazione si può dedurre che, in realtà, attraverso l'analisi di k-Means vengono definite condizioni meteomarine più lievi, sia per stati ambientali molto intensi sia per stati di intensità ridotta. In effetti, gli stati ottenuti attraverso il k-means si riferiscono alle condizioni medie dei dati appartenenti ad un particolare cluster, mentre gli stati selezionati da MDA si avvicinano ai bordi degli scatter nello spazio delle variabili. Per spiegare quanto affermato, un esempio è riportato in un caso 2D semplificato nella Fig. 18.

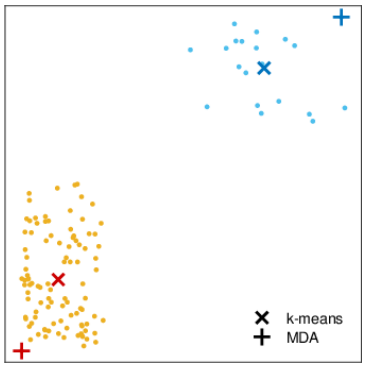

Figura 18. Confronto tra i centroidi calcolati con k-Means e con MDA.

Si vuole precisare che, quando viene applicata l'analisi tramite l'algoritmo di k-Means, il profilo dello stato target di H_s non corrisponde al centroide calcolato; infatti, si è scelto di selezionare lo stato più limitrofo al centroide e non il centroide stesso. Questo permette di considerare veri e propri stati reali del database hindcast in esame come rappresentanti di ogni cluster, e non punti fittizi costruiti calcolando la media di più variabili. D'altro canto, il profilo dello stato target di H_s individuato con MDA non ha bisogno di ulteriori modifiche dal momento che l'algoritmo non prevede il calcolo della media.

Tali considerazioni sono da ritenersi valide anche per l'analisi regionale. Nelle figure 13-16 si osservano gli andamenti delle altezze d'onda in corrispondenza degli eventi di libeccio e di tramontana ottenuti per ogni punto del sottobacino. I risultati permettono di vedere a confronto come variano nello spazio geografico le condizioni climatologiche a parità di stato significativo considerato. È infatti possibile notare come varia la direzione di propagazione delle onde nei diversi punti analizzati in funzione della loro posizione geografica.

Infine, la Fig. 12 suggerisce un'altra considerazione. Si può notare che la serie di θ_p non appartiene solo al settore della tramontana durante l'intera settimana; infatti, sono presenti alcune onde che si trasformano spostandosi nel settore dello scirocco (direzione in entrata SE) e del libeccio. Quanto evidenziato è una caratteristica presente in tutti gli stati ottenuti, tranne quelli relativi a intensi eventi libeccio. In effetti, gli stati meteomarini meno acuti non mostrano caratteristiche uni-modali durante lunghi periodi.

Come è noto, la morfologia della costa ligure la rende più esposto a eventi di libeccio (e questo è dovuto al fetch dell'area in esame): infatti, è possibile rilevare che l'area in esame è caratterizzata da intere settimane determinate da onde in arrivo a sud-ovest. Le mareggiate di tramontana e di scirocco non mostrano caratteristiche così intense, e questo tipicamente si riflette sul fatto che tali eventi di solito sono più brevi di una settimana.

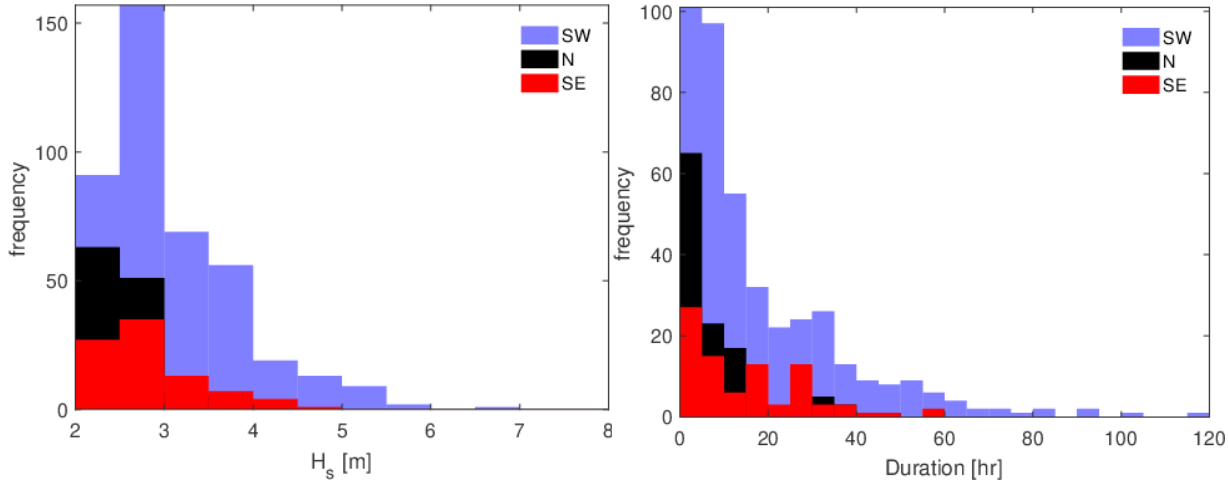

Figura 19. Analisi di altezza e durata delle mareggiate nell'area investigata.

Ciò può essere apprezzato osservando la Fig. 20, in cui viene mostrata la distribuzione della frequenza di H_s e la durata delle mareggiate marine effettuate nel Punto 000230 (vedasi Fig. 1). Le tempeste sono individuate attraverso un'analisi Peak Over Threshold (di solito denominata "POT"): è uno dei principali metodi per l'Analisi ai Valori Estremi e si basa sull'estrazione, da una registrazione continua, dei valori di picco raggiunti in un periodo, che si realizzano oltre una soglia. Vengono prese, quindi, in esame tutte le altezze d'onda che superano una certa soglia, impostata pari al 98% quantile della distribuzione iniziale di H_s. Quindi, i superamenti di H_s sono stati suddivisi considerando un intervallo di tempo minimo tra gruppi differenti di almeno 24 ore (infatti due gruppi di altezze d'onda sopra soglia sono considerati eventi separati solo se si verificano a distanza di almeno un giorno). Infine, una volta che gli eventi sono stati definiti, viene mantenuto il massimo di altezza d'onda e viene calcolata la durata totale di ciascun gruppo di superamenti. Come mostra la Figura 20, gli stati marini più intensi possono durare fino a 5 giorni per il settore di libeccio, mentre negli altri casi si esauriscono al massimo in un paio di giorni (riquadro destro) e sono caratterizzati da valori inferiori di H_s (pannello di sinistra).

5 Conclusioni

In questo lavoro, è stata sviluppata una metodologia capace di selezionare degli scenari climatologici sulla base di un dataset costituito da variabili della dinamica della circolazione

marittima. L'obiettivo infatti è quello di ridurre i tempi di calcolo dei modelli di Fluidodinamica Computazionale, concentrando l'attenzione solo su alcune condizioni meteomarine significative che vengono successivamente risolte numericamente. In questo caso, il fenomeno analizzato riguarda lo studio della dispersione delle microplastiche nelle acque portuali e zone circostanti.

Tale metodologia è stata sviluppata introducendo l'applicazione di tecniche di "Data mining" tramite algoritmi di clusterizzazione: infatti il clustering è una tecnica di analisi multivariata e permette creare dei gruppi di elementi, a partire da dataset di grandi dimensioni, sulla base della loro "lontananza logica". In particolare, sono stati impiegati il k-Means e MDA che risultano essere due degli algoritmi più diffusi in questo ambito.

L'applicazione di questi algoritmi è stata inizialmente effettuata in corrispondenza di un singolo punto situato al largo del porto di Genova e ha permesso di individuare 30 scenari climatologici: nel caso del k-Means si tratta di stati che esprimono la variabilità media del dataset di partenza, mentre nel caso di MDA si ottengono stati rappresentativi anche di condizioni estreme.

Successivamente, è stata applicata la clusterizzazione regionale che ha permesso di individuare 30 cluster con i quali è possibile descrivere in dettaglio il clima marino del sottobacino e la sua variabilità: ciascuno stato risulta rappresentativo di una particolare condizione climatologica per ogni punto geografico della zona.

I risultati ottenuti sono stati quindi validati confrontandoli con la climatologia media del sottobacino in esame. La validazione porta a concludere che l'implementazione di tale metodologia può diventare uno strumento utile per la definizione di scenari multivariati permettendo una importante riduzione dello sforzo computazionale dei modelli CFD e una significativa ottimizzazione nell'impostazione del lavoro effettuando simulazioni numeriche più mirate.

1 Introduction

L'analyse des processus physiques liés à la circulation côtière et aux processus de dispersion dans cette zone a toujours été d'un grand intérêt, tant du point de vue scientifique qu'applicatif. La nécessité de disposer d'outils d'analyse et de prévision de ce type de processus se fait sentir non seulement dans le domaine de la planification et de la programmation des activités humaines dans la bande côtière (c'est-à-dire le contrôle de la dispersion du dragage, la conception des sorties des stations d'épuration), mais aussi dans le domaine de la gestion des situations d'urgence et de la qualité de l'eau (c'est-à-dire les déversements et la dispersion des polluants tant du côté terrestre que du côté maritime, les accidents en mer et les opérations de recherche et de sauvetage). De ce point de vue, au cours des dernières décennies, notamment grâce à la croissance exponentielle de la puissance de calcul des ordinateurs modernes, on s'est de plus en plus appuyé sur l'utilisation de programmes de calcul capables d'effectuer des simulations numériques de l'hydrodynamique marine et côtière et des processus de dispersion tant du point de vue eulérien (dynamique de concentration) que lagrangien (masse et dispersion des objets). En particulier, un large éventail de modèles numériques a été développé pour l'analyse, la simulation et la résolution des problèmes de dynamique des fluides : il s'agit des modèles de dynamique des fluides numérique (CFD) qui permettent de simuler les phénomènes environnementaux les plus importants en résolvant numériquement les lois physiques des fluides. En fait, la CFD est principalement utilisée pour résoudre numériquement des équations dans des situations réelles et complexes.

Il est ainsi possible, par exemple, d'analyser les processus morphodynamiques, en étudiant l'évolution du littoral à la suite de tempêtes intenses et d'éventuels phénomènes d'érosion. Le CFD permet d'étudier la circulation marine afin d'approfondir la connaissance des courants qui caractérisent une zone donnée et de traiter les problèmes potentiels liés à la dispersion des polluants. Il existe également des modèles, en phase de conception, qui permettent d'évaluer le comportement des structures côtières en fonction de conditions de mer particulières.

L'étude de tels phénomènes repose généralement sur une énorme quantité d'informations qui nécessitent des temps de calcul élevés et de grandes puissances de calcul qui ne sont pas toujours disponibles : cela peut notamment se produire lorsque, par exemple, le jeu de données examiné provient d'un service de réanalyse climatologique, caractérisé par une haute résolution temporelle et spatiale. Dans ce cas, il peut être opportun de réduire le nombre de conditions environnementales à prendre en compte pour les simulations numériques afin d'identifier et de préserver les modes de variabilité les plus significatifs du phénomène. La résolution d'un nombre limité de conditions environnementales, également appelées "scénarios", est avantageuse non seulement parce qu'elle permet de sélectionner les scénarios les plus importants pour le processus étudié, mais aussi parce qu'elle réduit considérablement la charge de calcul nécessaire

pour résoudre l'ensemble de la chaîne de modélisation.

À cette fin, il est possible d'utiliser des techniques de "Data mining", c'est-à-dire d'analyse massive des données disponibles, grâce à des algorithmes de regroupement ; cette approche s'est révélée très utile : elle permet en effet de regrouper un ensemble de données en classes d'objets (clusters) sur la base de leur similarité/dissimilarité. Un groupe représente un regroupement d'éléments qui sont similaires les uns aux autres et qui sont différents des éléments d'un autre groupe. Le résultat est un sous-ensemble d'éléments qui peuvent résumer l'ensemble de données initial tout en conservant ses principales propriétés.

Dans la littérature scientifique, il existe plusieurs applications des techniques de regroupement pour l'identification des conditions environnementales avec différents objectifs spécifiques qui concernent non seulement l'analyse des processus physiques mais aussi leur simulation numérique. L'application de tels algorithmes aux bases de données de réanalyse des conditions météorologiques et marines (généralement extrêmement importantes en termes de nombre de variables stockées, en particulier dans le cas d'une haute résolution spatiale et temporelle) permet de décrire les conditions météorologiques et marines, en sélectionnant certains états représentatifs de sa variabilité, dans le but de les mettre en œuvre dans une méthodologie de propagation du mouvement des vagues (Camus, Mendez, Medina, & Cofiño, 2011b). Pour la définition des états de mer, les auteurs considèrent des séries chronologiques horaires de la hauteur, de la période et de la direction moyenne des vagues, sans tenir compte de leur évolution dans le temps.

Différente est l'approche de (Bárcena, Camus, García, & Álvarez, 2015) qui utilise des techniques de clustering dans la simulation de l'hydrodynamique tridimensionnelle des estuaires avec une haute résolution spatiale et temporelle : en effet, une fenêtre temporelle est initialement définie sur la base de laquelle on analyse les données de départ, afin d'obtenir des clusters représentés par des séries temporelles définies. L'utilisation de cette approche permet donc de représenter synthétiquement la variabilité des marées astronomiques et d'identifier des scénarios des forces de forçage en jeu pour obtenir le comportement moyen réduisant la taille de l'ensemble de données initial.

Il apparaît que le choix de la longueur de la fenêtre temporelle des données à examiner dépend non seulement de l'échelle de temps des forçages considérés mais aussi du temps caractéristique du processus à examiner ultérieurement. Par exemple, la propagation du mouvement des vagues du large vers la côte est étudiée en considérant les états de mer horaires, alors que l'échelle de temps typique des problèmes de dispersion est de l'ordre de quelques semaines.

Une autre particularité des techniques décrites dans les travaux mentionnés ci-dessus est la

manière dont la mise en grappes est appliquée : d'une part (Camus, Mendez, Medina, & Cofiño, 2011b) regroupent les variables en les considérant de manière conjointe, d'autre part (Bárcena, Camus, García, & Álvarez, 2015) effectuent une mise en grappes pour chaque variable impliquée dans l'analyse de manière indépendante.

Cependant, l'étude des processus dans le domaine de la météorologie ne permet pas de considérer les variables indépendamment les unes des autres : dans ce travail, en effet, nous proposons une méthodologie qui permet de caractériser le climat marin en considérant non seulement les caractéristiques du mouvement des vagues mais aussi la vitesse du vent, le champ de pression et le forçage des marées. Pour la construction de l'ensemble de données initial, une fenêtre temporelle appropriée est donc choisie en fonction du type de processus à étudier, c'est-à-dire la description de la dispersion des polluants/sédiments/particules dans les eaux côtières, suite au rejet en mer d'un débit défini dans un certain intervalle de temps.

2 Données et Méthodes

2.1 Paramètres Météo-Océaniques

Les variables météo-marines utilisées dans cette étude proviennent des produits rétrospectifs du Département d'ingénierie civile, chimique et environnementale de l'Université de Gênes (DICCA), www3.dicca.unige.it/meteocean/hindcast.html. Grâce à une nouvelle analyse des conditions météorologiques, une base de données a été construite contenant des données horaires de vagues, de vent et de champ barique définies sur une grille avec une résolution d'environ 10 km lon/lat, étendue à l'ensemble du bassin de la mer Méditerranée (Mentaschi, Besio, Cassola, & Mazzino, Developing and validating a forecast/hindcast system for the Mediterranean Sea., 2013 ; Mentaschi, Besio, Cassola, & Mazzino, Performance evaluation of wavewatch iii in the mediterranean sea., 2015). La mise en œuvre de l'ensemble de données rétrospectives a eu lieu après la validation et l'optimisation de la chaîne de modèles numériques utilisée (WRF pour la partie météorologique et WaveWatchIII pour la partie sur les vagues) et, à ce jour, ces données ont été utilisées dans de nombreuses recherches et applications (Re, Manno, Ciraolo, & Besio, 2019), (Leo, Besio, Zolezzi, & Bezzi, 2019), (Sartini, Besio, Dentale, & Reale, 2016), (De Girolamo, et al, 2018), (Sartini, Besio, & Cassola, 2017), (Zughayar, Gudmestad, De Leo, & Besio, 2017), (Mucerino, et al., 2019), (Besio, Briganti, Romano, Mentaschi, & Girolamo, 2017), (Ferretti, et al., 2018). Les séries chronologiques de 1979 à 2018, sur une base horaire, de la hauteur significative des vagues (H_s), de la période et de la direction des pics (T_p et θ_p, respectivement) des composantes de la vitesse longitudinale et latitudinale du vent (w_x/w_y) et de la pression au niveau moyen de la mer (mslp) en un point de grille devant Gênes (mer Tyrrhénienne, voir Fig. 1) sont prises en compte pour le développement des algorithmes d'identification des scénarios climatiques caractéristiques. Ensuite, les forces de forçage de la marée (ci-après $\Delta\eta$) sont

obtenues à la position sélectionnée à l'aide du logiciel de prévision des marées (TPXO.3) fourni par l'université de l'État de l'Oregon (Egbert & Erofeeva, 2002). L'excursion de la marée a été calculée dans le même intervalle de temps et avec la même fréquence pour lesquels des données météorologiques rétrospectives étaient disponibles.

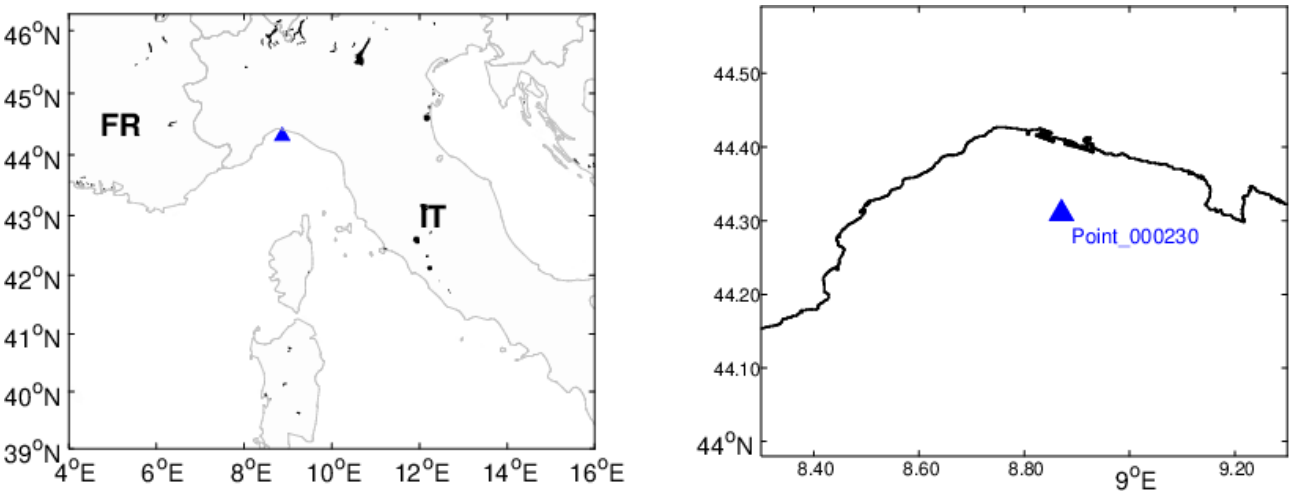

Figura 2: Zoom sur la zone d'étude. Le point sélectionné est mis en évidence par un triangle bleu accompagné du code numérique du point rétrospectif (lon/lat-8.8707/44.31 ; système de référence : WGS 84).

2.2 Analyse des Clusters

Le clustering est un ensemble de techniques d'analyse de données multivariées visant à sélectionner et à regrouper des éléments homogènes dans un ensemble de données, sur la base de mesures de similarité entre les éléments eux-mêmes, en termes de distance dans un espace multidimensionnel. Il peut être utilisé pour examiner les distributions de données, pour observer les caractéristiques de chaque distribution et pour se concentrer sur celles qui présentent le plus d'intérêt. Il peut également être utilisé comme étape de prétraitement des données pour d'autres algorithmes fonctionnant sur les grappes identifiées (comme dans le cas présent).

Le point de départ des techniques de cluster expliquées ci-dessous est la modification de l'ensemble de données original pour la construction d'une matrice de données $X_{n,V}$ où n et V sont respectivement le nombre de données à modéliser et le nombre de variables problèmes. En prenant V comme vecteurs de n données, $X_{n,V}$ est défini comme :

$$X_{n,V} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,V} \\ \vdots & \ddots & \vdots \\ x_{1,n} & \cdots & x_{n,V} \end{bmatrix} \quad (1)$$

Le regroupement des données (ou "clustering") à partir de l'ensemble de données ainsi défini permet de sélectionner un certain nombre de sous-ensembles significatifs, constitués de lignes

de la matrice X, quelle que soit la structure temporelle des variables elles-mêmes. Cette approche permet de décrire de manière exhaustive la variance des forçages à travers un nombre limité d'états : en effet, l'utilisation de l'ensemble des données peut parfois devenir gênante non seulement d'un point de vue informatique (dans le cas de trop nombreuses simulations de conditions météo-marines) mais aussi du point de vue de la caractérisation d'un scénario type [voir (Camus, Mendez F. J., & Medina, 2011a) pour certaines applications].

Dans cet ouvrage, il est fait référence aux techniques de classification par catégories qui permettent de classer les données puisque l'étiquette de classe a priori n'est pas connue. L'algorithme dit de dissimilarité maximale (désormais MDA) et le k-Means, deux des algorithmes les plus populaires dans le domaine de l'analyse par fouille de données, sont pris en considération : ils permettent de partitionner les données sur la base de leur dissimilarité/similarité et permettent de choisir quelles sont les caractéristiques d'intérêt pour distinguer les différents groupes (en fonction du choix du nombre de clusters).

2.2.1 MDA

L'objectif de MDA est de sélectionner un sous-ensemble de la matrice $X_{n,V}$, définie $X^*_{m,V}$ (matrice d'étiquettes, où $m < n$), qui représente le mieux la variance globale des données. L'étape initiale consiste à normaliser toutes les variables scalaires (c'est-à-dire chaque colonne de la matrice $X_{n,V}$) dans un espace commun, afin de pouvoir travailler facilement avec des variables caractérisées par plusieurs ordres de grandeur. L'algorithme de sélection commence par l'identification des premières données significatives, identifiées comme l'état le plus éloigné du nuage de points initial, c'est-à-dire celui qui est le plus "dissemblable" du reste de l'ensemble de données. Une fois sélectionné, on procède au calcul de la dissimilitude entre les vecteurs de la matrice $X_{n-1,V}$ et du sous-ensemble $X^*_{1,V}$, puis au calcul de la distance euclidienne ainsi exprimée :

$$d_{i1} = \|x_i - x_1^*\|, \;\; i = 1, \dots, n-1 \qquad (2)$$

où x e x^* indiquent respectivement les lignes de la matrice d'entrée et de sortie. Le nouvel élément est alors sélectionné comme celui caractérisé par la valeur maximale de d_{i1} et ajouté à la matrice cible X^*. Comme illustré dans (Camus, Mendez, Medina, & Cofiño, 2011b), la version MaxMin de l'algorithme est appliquée : en effet, en correspondance de la k^{ième} itération ($k < m$, où k est le nombre d'éléments de la matrice X^*), la distance à considérer pour chaque élément de $X_{n-k,V}$ est le minimum par rapport à tous les k vecteurs de $X^*_{k,V}$; parmi toutes ces distances, le maximum est sélectionné et l'état correspondant est par conséquent ajouté à X^*. Le calcul se termine lorsque k est égal à m, c'est-à-dire une fois que le nombre de grappes (préalablement établi) est atteint. À la fin du processus de construction de la matrice X^*, on obtient alors un sous-ensemble de vecteurs qui permet de résumer l'ensemble de données de départ avec un nombre

d'états plus petit : les éléments restants sont affectés au vecteur modèle le plus proche correspondant, formant ainsi de véritables classes d'états.

2.2.2 k-Means

Le k-Means est un algorithme de regroupement partiel qui permet de diviser un ensemble d'objets (ou de vecteurs) en k groupes, en fonction de leurs attributs. Il s'agit d'une technique basée sur le calcul de la distance euclidienne entre les différents éléments de l'ensemble de données, comme dans le cas de la MDA. Toutefois, dans ce cas, l'objectif est de minimiser la variance intra-groupe, chacun étant identifié par un centroïde ou un point médian de la même taille que les données originales (MacQueen, 1967). Partant à nouveau d'une normalisation des variables, l'algorithme suit une procédure itérative en attribuant, à la première étape, les centroïdes de façon aléatoire (c'est-à-dire $X_{m,V}^{*,1}$), choisis parmi les lignes de la matrice $X_{n,V}$. Ensuite, chaque donnée (par exemple, la première ligne de $X_{i,V}$, $i \in [1,n]$) est "assignée" au centroïde le plus proche :

$$m_i = m/\left(d_i = \|x_i - x_{m,V}^{*,1}\|, i = 1, \dots n - m\right) \quad (3)$$

où m_i est le premier groupe auquel les données appartiennent. Une fois que les m groupes ont été définis, les nouveaux centroïdes ($x_{m,V}^{*,2}$) sont calculés simplement comme une moyenne des groupes respectifs :

$$x_{m,V}^{*,2} = \sum_{x_i \in m_j} \frac{x_i}{n_j} \quad (4)$$

n_j étant le nombre d'éléments appartenant au jème groupe. Le processus de classification se termine lorsque la position des centroïdes ne change pas de manière significative entre deux itérations successives : dans ce cas, on dit que l'algorithme a atteint la convergence.

2.2.3 Sélectionner le nombre optimal de clusters

Les techniques de clustering décrites ci-dessus nécessitent comme première étape fondamentale de définir un nombre approprié de clusters. Le choix peut être fait subjectivement, si l'utilisateur, par exemple, veut que les données soient affectées à un certain nombre de classes, ou s'il sait comment et avec quelles distributions les données sont présentées (comme dans l'exemple de la section 2.5). Cependant, si le nombre optimal de grappes n'est pas connu a priori, il est nécessaire d'introduire une analyse de sensibilité des résultats de la mise en grappes. Par exemple, (Bárcena, Camus, García, & Álvarez, 2015) fait référence à l'indice CE (sous Model Efficiency) proposé par (Nash & Sutcliffe, 1970), qui permet d'évaluer l'efficacité des états du modèle dans la reproduction de l'ensemble de données de départ. Une approche similaire est appliquée dans (Núñez, et al., 2019) où le *Mean Skill Index* introduit par (Willmott, 1981) est

calculé : ce coefficient reflète la précision avec laquelle les variables classées se rapprochent des variables originales.

Dans ce travail, nous avons pris en considération ce qui est suggéré par (Solari, et al., 2017), c'est-à-dire l'utilisation de la "variance totale", calculée comme :

$$W^2 = \frac{1}{n}\sum_{i=1}^{m}\sum_{i=1}^{m_j}\left(x_j - x_i^*\right)^2 \quad (5)$$

n étant le nombre total d'états et m_j le nombre d'éléments appartenant au jème cluster ; x_i^* est le ième centroïde de la matrice X^*. L'équation (5) peut donc être utilisée, dans le cas de k-Means, comme un indicateur utile pour évaluer le nombre de clusters, conduisant à une valeur quasi-asymptotique de W^2.

Une approche similaire peut être appliquée dans le cadre du programme MDA, même si dans ce cas, le but du regroupement est d'expliquer la variance globale des données. Par conséquent, dans ce cas, il est fait référence au calcul de la distance moyenne entre les États modèles, définie comme :

$$\bar{d} = \frac{1}{m-1}\sum_{i=1}^{m-1} d_i \quad (6)$$

où d_i est la distance entre l'état cible i^{ème} et (i+1)^{ème} sélectionné par le processus itératif MDA. Par conséquent, m est défini comme la valeur au-delà de laquelle aucune modification significative de $\bar{d}$ ou W^2 n'est appréciée.

2.2.4 Clustering des données en fonction de l'évolution dans le temps

Jusqu'à présent, l'ensemble de données initial considéré pour l'application des algorithmes de regroupement (par exemple $X_{n,V}$), était défini simplement en flanquant la série temporelle originale de variables, chacune étant caractérisée par une résolution temporelle horaire. Chaque variable, en effet, est représentée par un vecteur et chacun de ses éléments décrit un état unique.

Cependant, l'objectif de ce travail est de sélectionner les états qui doivent être utilisés pour effectuer des simulations numériques de phénomènes physiques particuliers : il devient donc important de prendre en compte la variabilité et la tendance temporelle des forces en jeu. Il faut donc sélectionner des scénarios significatifs d'un point de vue temporel, ce qui nécessite de réorganiser les données d'entrée tout en conservant la structure temporelle des forces en question. Conformément à (Bárcena, Camus, García, & Álvarez, 2015), une fois que l'échelle de temps des forces a été définie, les ensembles de données respectifs peuvent être organisés en conséquence en séries d'extensions assignées, de manière à préserver la tendance des variables dans le temps. L'objectif d'une telle réorganisation des données est précisément d'essayer de

conserver l'empreinte des événements réellement observés. S'agissant du nombre de pas de temps (c'est-à-dire le nombre d'heures) dont se compose l'échelle de temps de référence, le vecteur x est redéfini de la manière suivante :

$$x'_{j,:} = x[i\delta + 1 : nt + i\delta], i = 0, \dots, (n - nt)/\delta, j = i + 1 \quad (7)$$

où δ représente le décalage entre deux intervalles de temps successifs (c'est-à-dire le nombre de pas de temps entre les points initiaux de deux x' successifs).

Par la suite, l'analyse de regroupement peut être effectuée directement sur la matrice réorganisée X' ; dans ce cas, les lignes de la matrice ne se réfèrent plus à des états ponctuels des différentes variables, mais sont au contraire des fenêtres temporelles pour chaque quantité individuelle examinée. Un exemple d'une telle opération effectuée sur la série chronologique d'une variable X est donné dans l'équation 8.

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{nt} \\ \vdots \\ x_n \end{bmatrix} \longrightarrow X' = \begin{bmatrix} x_1 & \dots & x_{nt} \\ x_{i\delta+1} & \dots & x_{nt+i\delta} \\ \vdots & \dots & \vdots \\ x_{n-nt+1} & \dots & x_n \end{bmatrix} \quad (8)$$

2.3 Choix de l'échelle de temps de référence

Comme le montre la section 2.2, la variabilité temporelle des forçages examinés joue un rôle fondamental dans la définition de la non-stabilité des scénarios du modèle. Il est en effet nécessaire de choisir un intervalle de temps représentatif, adapté à toutes les quantités prises en considération. Comme le confirment également (Bárcena, Camus, García et Álvarez, 2015), la longueur des séries à court terme déterminant le scénario du modèle est liée aux échelles de temps caractérisant le signal. En particulier, les auteurs établissent une longueur d'échelle pour les forçats de marée en fonction des cycles de marée typiques de la zone géographique examinée ; tandis qu'ils utilisent un indice d'estimation de la durée des impulsions de débit pour l'évaluation de l'échelle de temps du débit du fleuve.

Pour notre application, il a été décidé d'évaluer l'échelle de temps des différents paramètres en calculant leurs fonctions d'autocorrélation respectives (ACF) : cette fonction permet de définir le degré de dépendance entre les valeurs prises par une variable échantillonnée dans son domaine d'abscisses. En d'autres termes, l'ACF représente la corrélation croisée entre le signal à l'instant t et un autre instant placé à une certaine distance (retard), et permet donc de vérifier sa dépendance mutuelle. En fait, l'objectif est de déterminer la fréquence fondamentale du signal examiné. Étant donné un ensemble de données x de longueur n, la fonction d'autocorrélation pour un retard k donné est définie comme suit :

$$\begin{cases} ACF_k = \frac{c_k}{c_0} \\ c_k = \frac{1}{n}\sum_{t=1}^{n-k}(x_t - \bar{x})(x_{t+k} - \bar{x}) \end{cases} \quad (9)$$

avec c_k la covariance pour le retard k, c_0 la variance d'échantillon de la série historique, tandis que $\bar{x}$ est la valeur moyenne de la série. L'ACF d'une variable particulière permet d'évaluer l'échelle de temps sur laquelle, en moyenne, une variable fonctionne sur des états significatifs. L'ACF est évalué dans une limite de confiance (ou l'intervalle de confiance "IC") pour un niveau de signification donné α :

$$\begin{cases} IC_k = ACF_k \pm z_{\frac{\alpha}{2}} \times SE(ACF_k) \\ SE(ACF_k) = \sqrt{\frac{1+2\sum_{j=1}^{k-1} ACF_j^2}{n}} \end{cases} \quad (10)$$

où $z_{\alpha/2}$ est le quantile relatif à la plage $[\alpha/2, 1-\alpha/2]$ dans l'espace normal, $SE(ACF_k)$ représente l'erreur type estimée.

Pour chaque taille, la fonction d'autocorrélation est évaluée pour une fenêtre de dix jours, composée de 240 décalages : chaque décalage est lié à la résolution temporelle de l'ensemble de données, qui est dans ce cas égale à une heure. La fenêtre a traversé toute la période considérée (1979-2018), le délai initial étant décalé d'une heure à la fois. Enfin, les valeurs ACF résultantes ont été calculées en moyenne pour chaque décalage. α a été choisi égal à 5%. Les résultats de l'analyse qui vient d'être décrite sont présentés dans la figure 2.

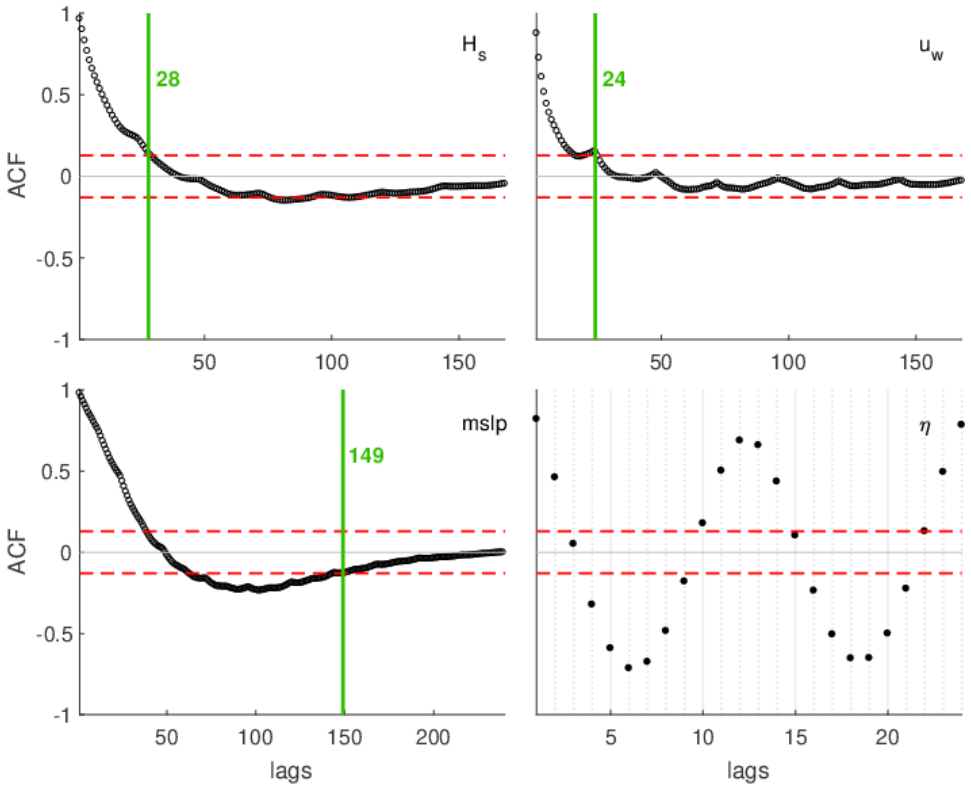

Figure 3: Fonction d'autocorrélation pour certaines variables météo-marines utilisées dans l'analyse.

Comme le montre la figure 2, l'échelle de temps pour les états de mer est d'environ un jour pour la hauteur de vague H_s et la vitesse du vent (se référant à u_w). Ce résultat reflète le climat marin local du point examiné (en face de Gênes), principalement caractérisé par des vagues de vent (les vagues dites de mer vive) : elles surviennent, se développent et disparaissent généralement à l'échelle quotidienne. Une considération similaire peut être appliquée à la période de vague T_p, qui est étroitement liée à H_s. D'autre part, la fonction d'autocorrélation pour la mslp nécessite près d'une semaine pour atteindre des valeurs comprises entre les limites respectives de l'intervalle de confiance (IC) : ceci est très probablement dû aux échelles synoptiques qui caractérisent les trajectoires des minima de pression. Enfin, comme on peut s'y attendre, l'oscillation de la surface libre est conditionnée par les marées de demi-journée et de quatrième quart, caractérisées par des cycles d'environ 12 et 6 heures respectivement.

Les résultats de l'ACF suggèrent que les variables impliquées dans l'analyse sont caractérisées à différentes échelles, il est donc nécessaire d'établir un schéma commun afin de faciliter l'application des analyses de regroupement ultérieures. Dans de telles situations, la fenêtre du modèle temporel (appelée Δt^*) doit être choisie en fonction du type de processus qui doit être simulé ultérieurement. Dans tous les cas, Δt^* ne doit pas être plus petit que la plus petite échelle de temps pour les variables considérées, afin que toutes puissent être caractérisées de manière adéquate sur leurs propres échelles représentatives. Sur la base de ces considérations et compte

tenu des temps d'échelle typiques des processus de dispersion, nous avons choisi, dans la présente étude, de fixer Δt^* à une semaine.

2.4 Analyse de corrélation entre les variables

Une fois l'échelle de temps de référence définie, il est nécessaire d'évaluer la corrélation qui caractérise les variables examinées. Comme le rapporte l'article de (Bárcena, Camus, García, & Álvarez, 2015), si les variables ne sont pas mutuellement corrélées, le regroupement peut être appliqué indépendamment pour chaque variable. En revanche, si la corrélation des variables considérées n'est pas négligeable, elles doivent être considérées ensemble.

Nous poursuivons donc l'évaluation des corrélations entre toutes les variables considérées ; en ce qui concerne les variables circulaires, (telles que θ_p et la direction du vent incident θ_w), nous introduisons l'indice de corrélation circulaire proposé par (Fisher & Lee, 1983) :

$$r = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} sinsin\,(\theta_i-\theta_j)\,sin(\alpha_i-\alpha_j)}{\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(\theta_i-\theta_j)\,sin^2(\alpha_i-\alpha_j)}} \tag{11}$$

Pour l'application de l'équation (10), seules quelques années de la série historique considérée ont été prises en compte : en particulier, seules trois années (à titre d'exemple) caractérisées par des intensités de vague différentes ont été prises en compte, afin d'avoir une idée de la tendance différente de r. Pour chaque année analysée, une hauteur de vague seuil (H_{th}) a été fixée et seule la direction de la vague de pointe θ_p et la direction de vent θ_w liées aux hauteurs ont été conservées, de sorte que la condition suivante $H_s \geq H_{th}$ se produit. Les résultats obtenus par l'analyse de corrélation sont présentés dans les figures 3 et 4.

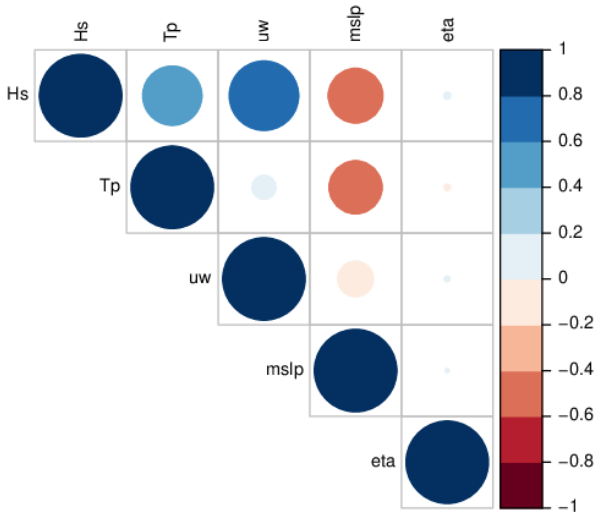

Figure 4: Corrélation des tailles non directionnelles.

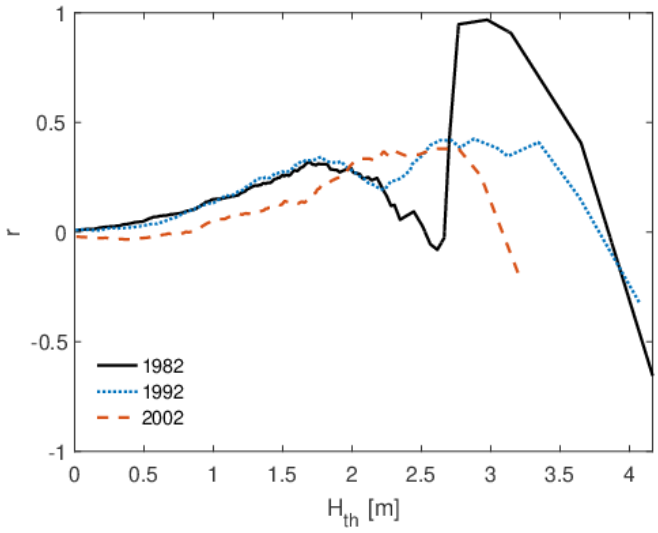

Figure 5 : Corrélation circulaire tra θ_p et θ_w pour différentes années et hauteurs de vagues seuils.

L'analyse de la figure 3 montre que H_s, T_p et u_w sont fortement corrélés, alors qu'ils sont anti-corrélés avec les mslp. En fait, les systèmes de basse pression sont associés à des états de mer intenses, qui à leur tour sont liés à des vitesses de vent élevées. Cela peut également être déduit des valeurs de corrélation circulaire entre les directions des vagues et du vent ; comme le montre la figure 4, r atteint des valeurs significatives en correspondance avec les conditions de houle, tandis que pour les états de mer extrêmes, il s'annule jusqu'à devenir négatif, ce qui indique une anti-corrélation entre les directions de la mer et du vent incident. Cela est probablement dû au fait que les états de mer extrêmement intenses ne sont pas dus au vent, mais sont liés à de longues vagues, générées loin de la zone d'échantillonnage et se déplaçant sur de longues distances (ce qu'on appelle les "swell").

Enfin, en ce qui concerne $\Delta\eta$, il n'y a pas de preuve de corrélation significative avec les autres variables considérées, puisque les cycles de marées astronomiques n'influencent pas le climat des vagues et ne dépendent pas de la pression moyenne au niveau de la mer. Par conséquent, les oscillations de la surface libre pourraient être regroupées indépendamment des autres paramètres qui, au contraire, doivent être traités comme un tout.

2.5 Clustering des variables circulaires

Les techniques de regroupement des données, expliquées dans la section 2.2, impliquent le calcul des distances euclidiennes entre les éléments d'un ensemble de données et certains éléments cibles définis dans le même espace, dans le but de regrouper les données selon l'algorithme utilisé.

Toutefois, lors de l'utilisation de variables circulaires (telles que la direction de la propagation des ondes), une précaution doit être prise pour permettre une manipulation aisée de ces données et

pour éviter les erreurs en présence de discontinuités dans l'espace des variables. Pour mieux expliquer ce concept, on peut se référer à θ_p, défini conformément à la convention nautique (les directions sont définies dans le sens des aiguilles d'une montre, en partant du Nord). Par exemple, on peut considérer deux ondes venant du Nord caractérisées par des directions d'arrivée de 0° et 360° respectivement : dans ce cas, la première est plus proche (c'est-à-dire plus proche) d'une onde cible se propageant vers l'ouest, tandis que la seconde est plus proche d'une onde orientée vers l'est (les θ_p respectifs étant égaux à 90° et 270° respectivement). Néanmoins, les deux conditions de vagues considérées ont la même direction d'origine et pour cette raison, un groupement différent serait insensé. Pour résoudre ce problème, une correction est généralement appliquée directement aux directions. Définissez θ_1 et θ_2 comme étant les directions des deux vagues examinées :

$$\Delta\theta = \begin{cases} 2\pi - (\theta_1 - \theta_2), & (\theta_1 - \theta_2) > \pi \\ (\theta_1 - \theta_2), & (\theta_1 - \theta_2) \in [-\pi \div \pi] \\ (\theta_1 - \theta_2) + 2\pi, & (\theta_1 - \theta_2) < -\pi \end{cases} \quad (12)$$

Toutefois, une approche plus appropriée est introduite et utilisée dans ce travail. Littéralement, toujours en utilisant les directions d'arrivée des vagues de l'exemple, θ_p peut être projeté le long de l'axe est-nord-est comme suit :

$$\begin{cases} \theta_{p,x} = -cos(\theta_p - 90) \\ \theta_{p,y} = sin(\theta_p - 90) \end{cases} \quad (13)$$

La correction appliquée aux arguments de la fonction sinusoïdale est nécessaire pour garantir que les composantes projetées sont conformes à la convention nautique : par exemple, les composantes sont toutes deux positives (négatives) dans le troisième (premier) quadrant, alors qu'elles sont de signes discordants dans le deuxième et le quatrième quadrant. À partir de l'équation 13, on peut déduire que $\theta_{p,x}$ et $\theta_{p,y}$ peuvent atteindre des valeurs négatives, qui sont évidemment sans signification d'un point de vue physique, mais qui sont significatives à des fins de regroupement, car elles permettent de regrouper les données en tenant compte à la fois des informations sur l'intensité des vagues et des directions d'origine.

La figure 5 montre un exemple synthétique d'un climat de vagues bimodal, composé de deux ensembles de données de vagues avec des directions de source distribuées autour de 0° et 180°, respectivement. L'algorithme k-Means a été appliqué en premier lieu en maintenant H_s et θ_p, puis en utilisant les projections dans le plan cartésien. Les résultats sont comparés dans la figure 6.

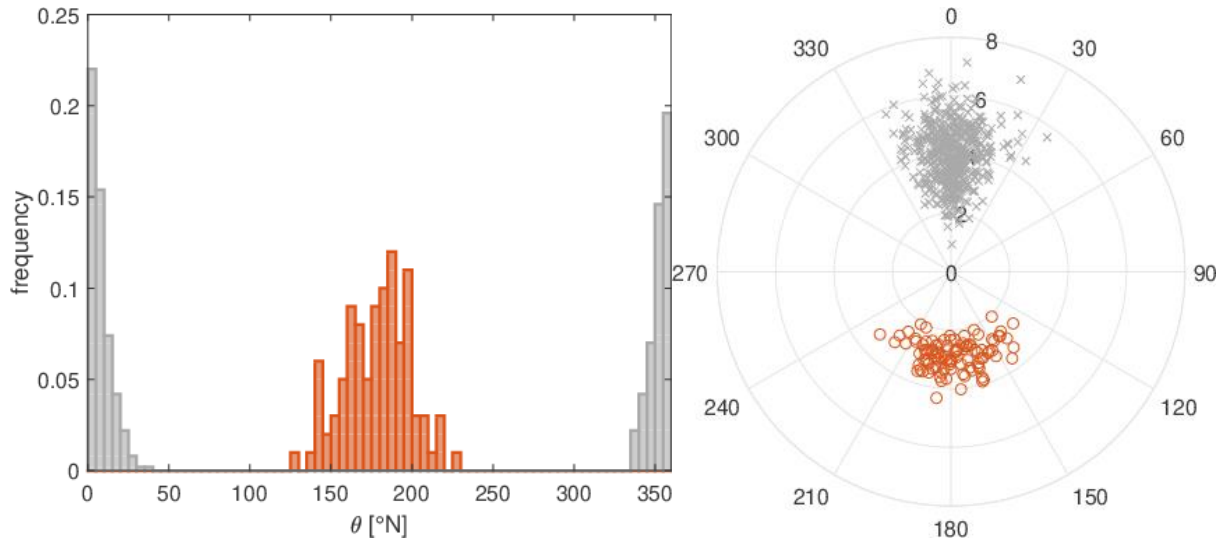

Figure 6 : Exemple d'un climat à ondes bimodales : distribution des fréquences (panneau de gauche) et graphique polaire (panneau de droite).

Lorsque le regroupement des états est effectué en tenant compte de θ_p e H_s séparément, et sans correction sur les distances circulaires, le regroupement des données est influencé par la direction d'origine des ondes, ce qui entraîne une classification incorrecte. Dans ce cas, les ondes sont réparties entre celles qui appartiennent aux quadrants 1-2 (croix grises) et aux quadrants 3-4 (cercles orange, encadré de gauche dans la figure 6). Au contraire, si l'on applique le clustering à $\theta_{p,x}$ et $\theta_{p,y}$, la classification est réussie : le caractère bimodal du climat des vagues est correctement détecté (panneau droit de la figure 6).

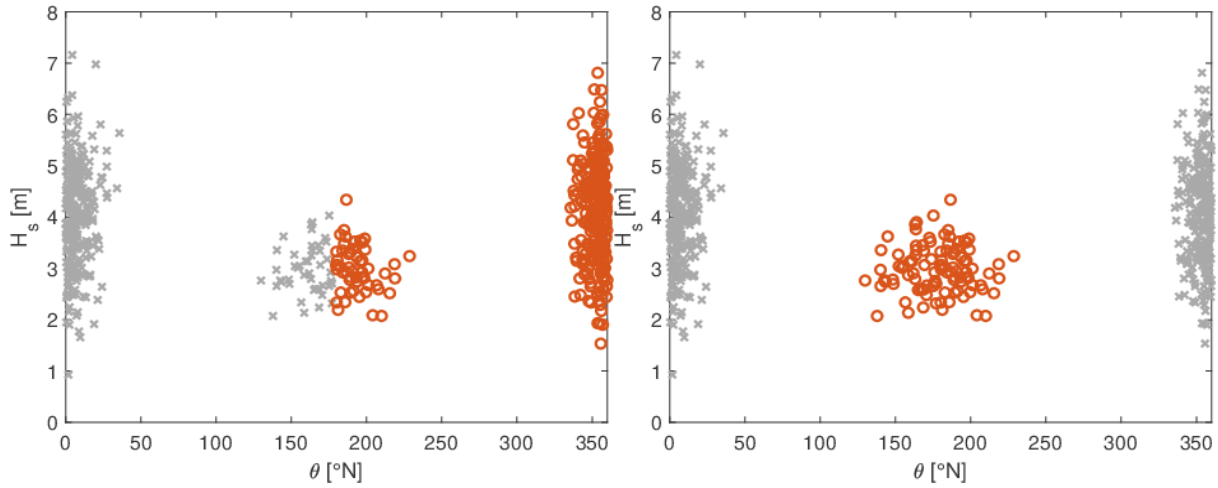

Figure 7 : Clustering de l'ensemble des données bimodales sur le mouvement des vagues comprenant : les variables originales (panneau de droite) et les projections (panneau de gauche).

2.6 Clustering régional

Parfois, les simulations numériques qui sont menées dans le domaine maritime ne se limitent pas à l'analyse d'un seul point, mais visent à étudier le comportement d'une zone géographique plus vaste (par exemple, une zone située devant un port ou un golfe). Il est donc proposé de poursuivre le développement de la méthodologie illustrée ici : c'est le clustering régional qui permet de prendre en compte la variabilité spatiale des paramètres météomar. L'objectif est d'obtenir une classification unique des conditions météo-marines d'une certaine zone géographique en considérant l'information étendue à un sous-bassin entier. De cette façon, il est possible de résumer avec précision le climat marin du sous-bassin en conservant quelques dizaines d'états qui sont en mesure d'exprimer la variabilité climatique de la zone examinée.

Le regroupement régional prévoit donc d'appliquer les algorithmes k-means et MDA à un ensemble de données qui contient les mêmes quantités de l'analyse précédente ($H_s, \theta_{p,x}, \theta_{p,y}, T_p, w_x, w_y, mslp, \Delta\eta$) de tous les points rétrospectifs appartenant au sous-bassin considéré. Reportez-vous à la carte de la Fig. 7 pour afficher l'emplacement des points analysés en rétrospective.

Figure 8 : Zoomez sur le sous-bassin qui vous intéresse. Les 7 points considérés sont marqués d'étoiles bleues.

3 Résultats

Conformément à l'analyse de la corrélation et de l'échelle de temps des variables, la matrice de données est construite en considérant la longueur de la fenêtre temporelle nt égale à 168 éléments (= 7*24 c'est-à-dire égale au nombre d'heures dans une semaine) pour chaque variable examinée : ensuite, les matrices x' de $H_s, \theta_{p,x}, \theta_{p,y}, T_p, w_x, w_y, mslp, \Delta\eta$ sont concaténées (voir Eq.7).

On obtient alors une matrice de dimensions égale à $V \times nt$, où V est le nombre de variables analysées (égal à 8).

La matrice X' a ensuite été normalisée le long de chaque colonne, en centrant les données autour de zéro (en soustrayant la moyenne) et en les mettant à l'échelle par rapport à l'écart-type :

$$X'_{:,j} = \frac{X'_{:,j} - \mu(X'_{:,j})}{\sigma(X'_{:,j})} \quad (14)$$

où μ et σ représentent respectivement la moyenne et l'écart-type de la j[ième] colonne de la matrice X'. En réalité, cette opération permet à l'algorithme appliqué ultérieurement de traiter à la fois les données négatives et positives, caractérisées par des ordres de grandeur différents, en évitant que les variables les plus pertinentes n'altèrent les calculs ultérieurs.

La figure 8 montre les résultats de l'analyse de sensibilité effectuée sur les données, tant pour l'algorithme MDA que pour l'algorithme k-Means (équations 5 et 6 respectivement). Les valeurs globales de W^2 et $\bar{d}$ ont été mises à l'échelle dans la plage 0-1, car l'échelle originale n'est pas pertinente, puisque les statistiques sont calculées à partir de variables normalisées ; la forme de la courbe est la seule caractéristique importante. Dans les deux cas, il est difficile de choisir un nombre exact de grappes, de sorte que le taux de variation pour $\bar{d}$ et W^2 est négligeable ; néanmoins, une pente plus douce à partir de 20/30 grappes est évidente. Ainsi, pour les calculs suivants, m est choisi égal à 30 pour les deux algorithmes considérés, afin de pouvoir comparer correctement les résultats des différentes techniques.

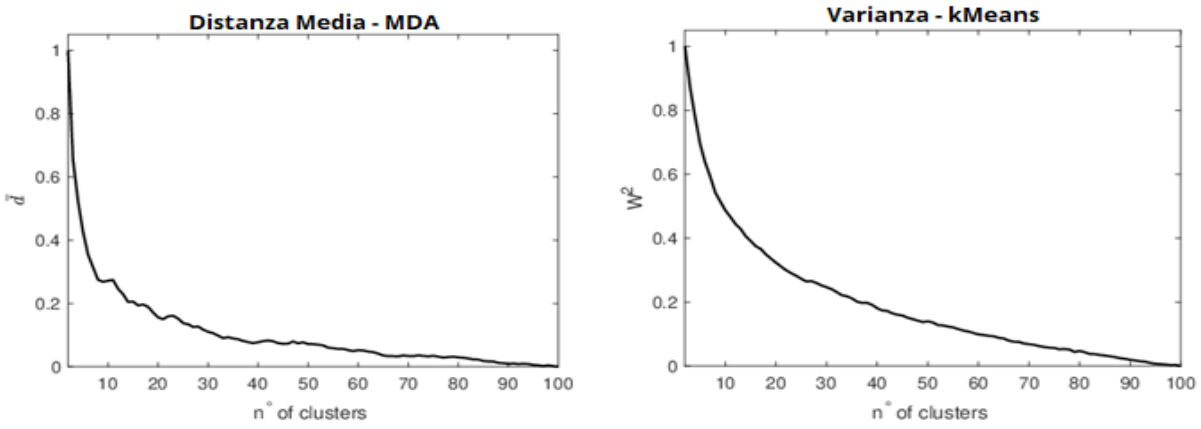

Figure 9 : Analyse de sensibilité en fonction du nombre de grappes ; algorithme MDA (panneau de gauche) et algorithme k-means (panneau de droite).

Les figures 9 à 12 ci-dessous montrent deux exemples (choisis parmi les trente groupes) d'états significatifs obtenus par l'application de k-Means et de MDA. Dans les figures 13 à 16, deux des trente groupes identifiés par regroupement de zones sont présentés à titre d'exemple. En

particulier, les états représentatifs d'un événement de libeccio et de tramontane résultant des deux analyses ont été choisis afin d'avoir une meilleure comparaison.

Il convient de noter que les directions θ_p et les hauteurs de vague significatives H_s des états cibles ont été obtenues en redéfinissant les variables $H_{s,x}$ et $H_{s,y}$, résultant des analyses de regroupement.

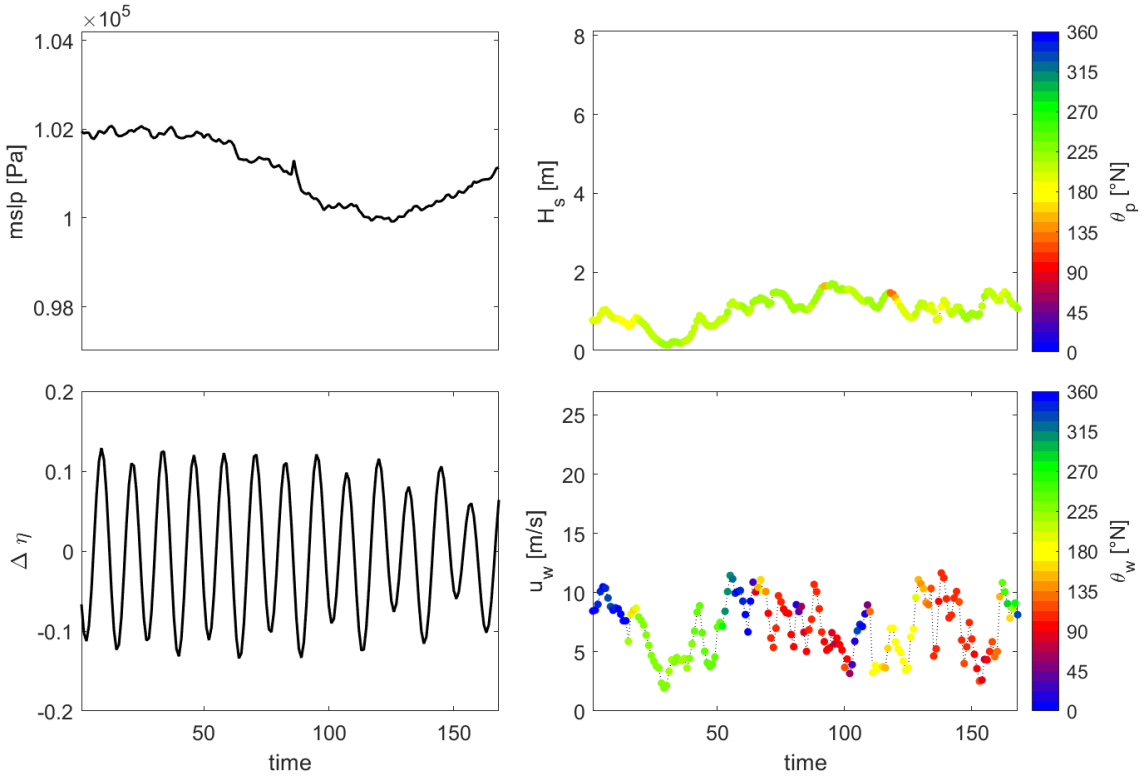

Figure 10: Statut significatif : caractéristiques d'un événement de libeccio obtenues par k-means.

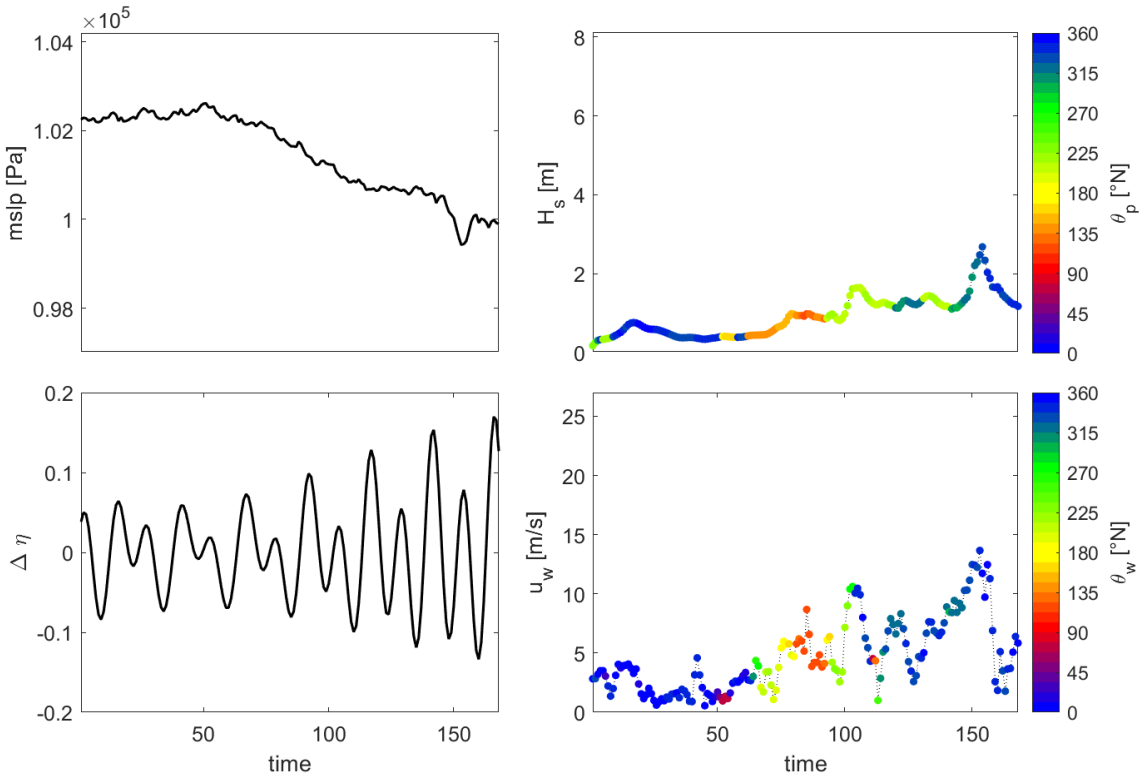

Figure 11 : État significatif : caractéristiques d'un événement mixte de la tramontane obtenues par k-means.

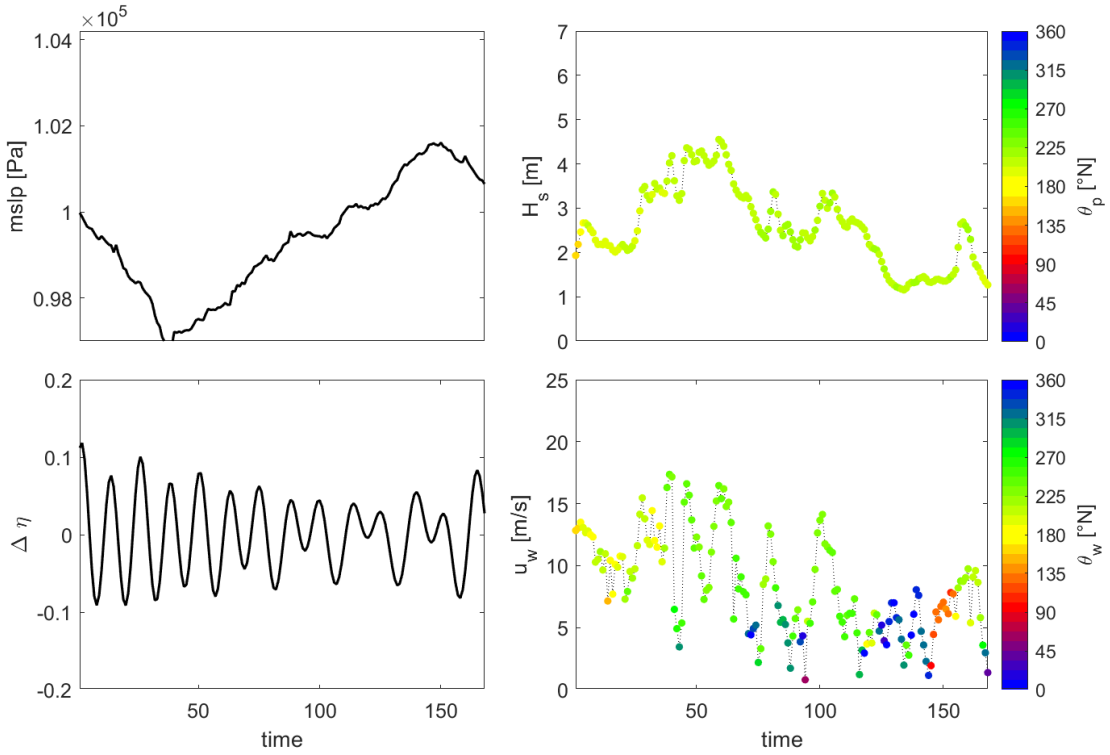

Figure 12 : Statut significatif : caractéristiques d'un événement de libeccio obtenues par MDA.

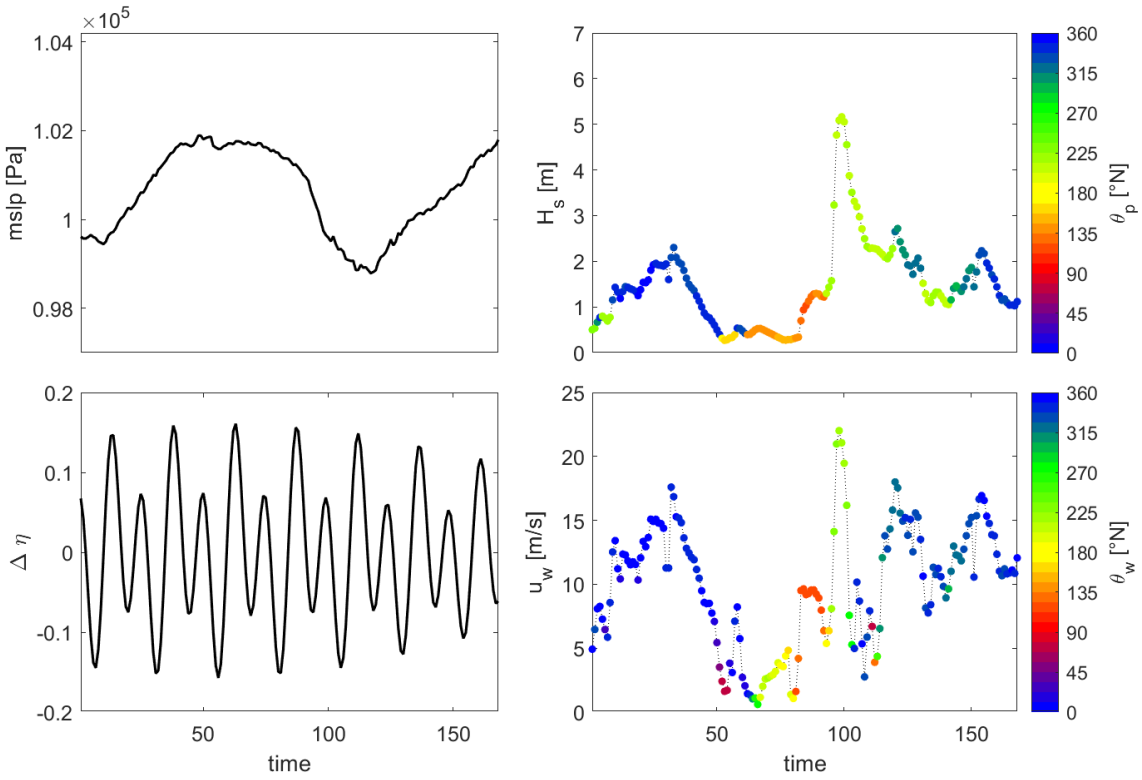

Figure 13 : État significatif : caractéristiques d'un événement mixte de la tramontane obtenues par MDA.

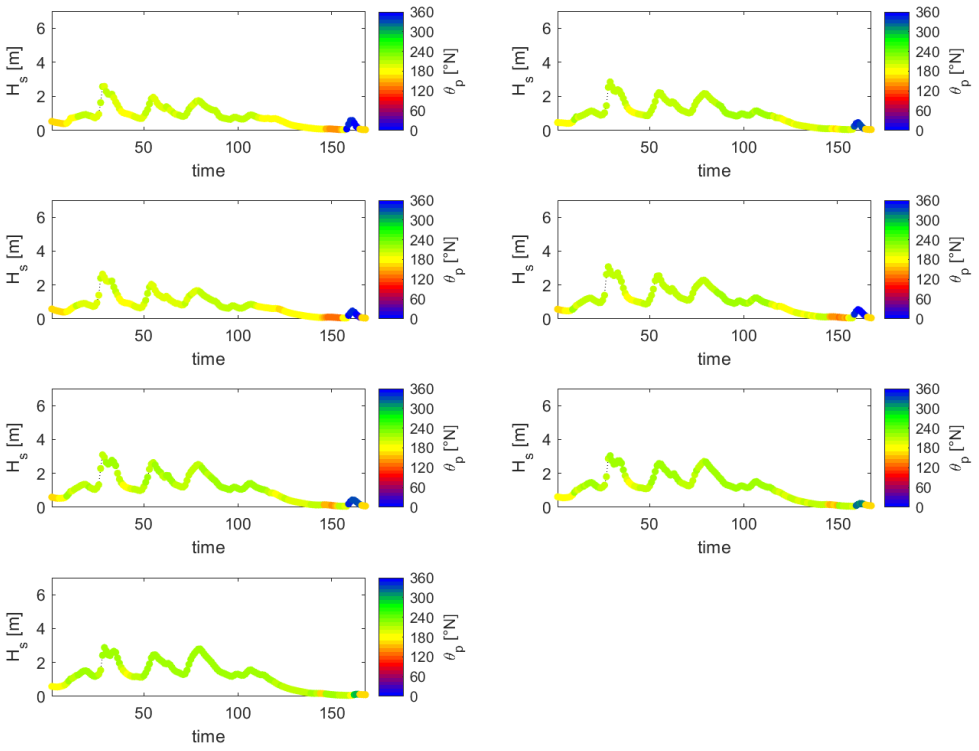

Figure 14 : Statut significatif : caractéristiques d'un événement de libeccio obtenues par k-means régional.

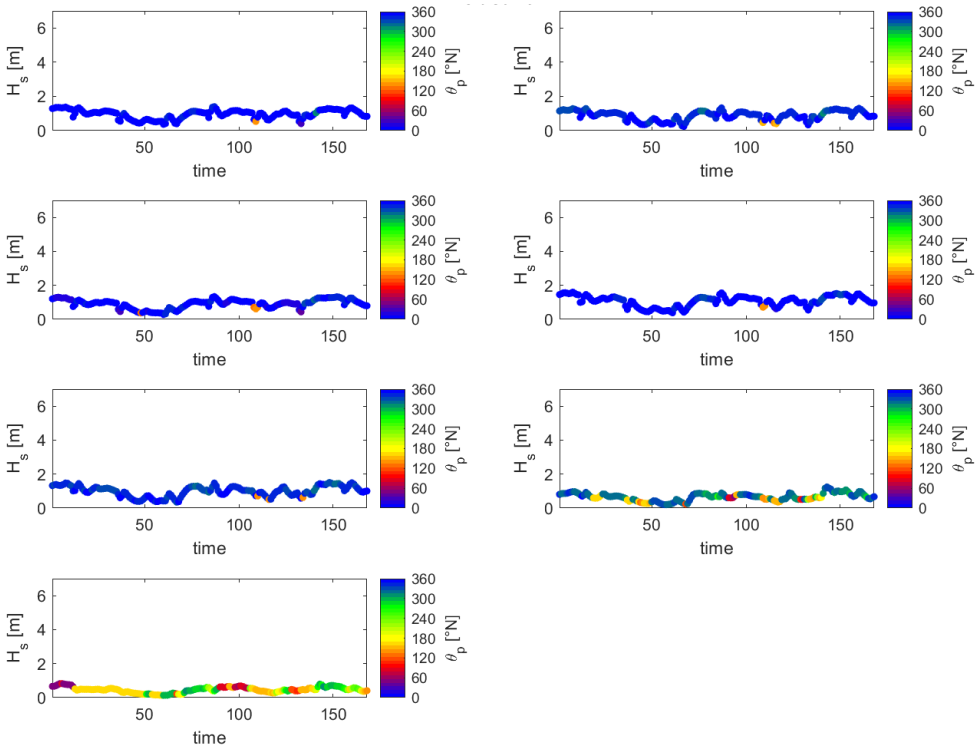

Figure 15 : État significatif : caractéristiques d'un événement de tramontane obtenues par k-means régional.

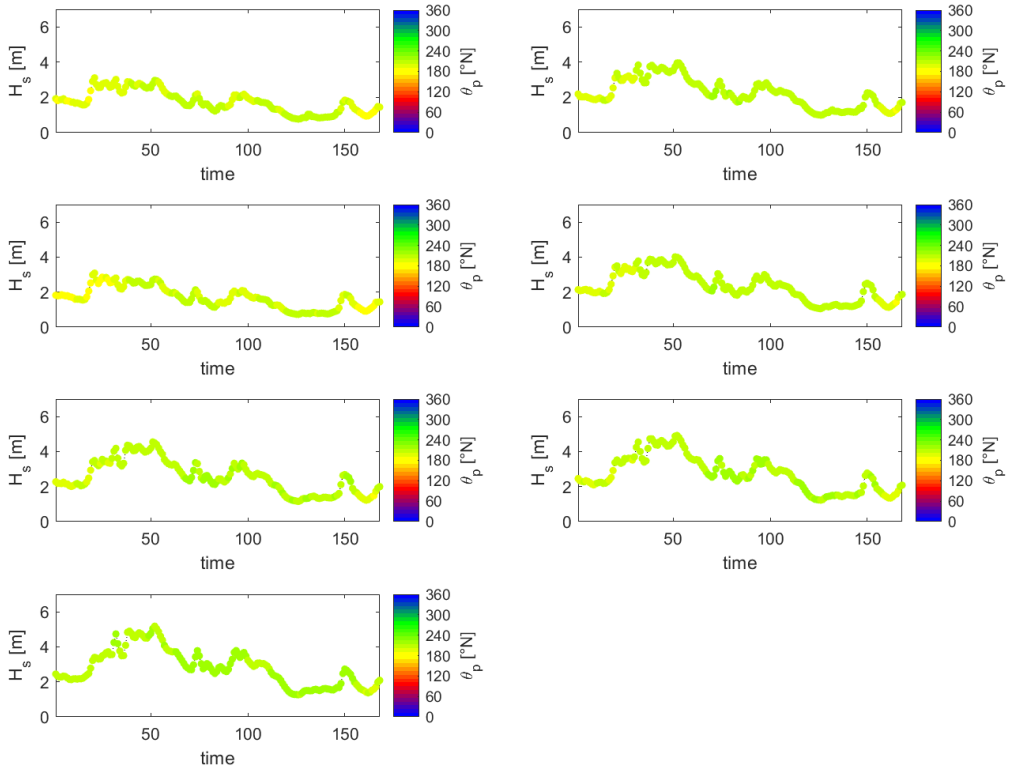

Figure 16 : Statut significatif: caractéristiques d'un événement de libeccio obtenues par MDA régional

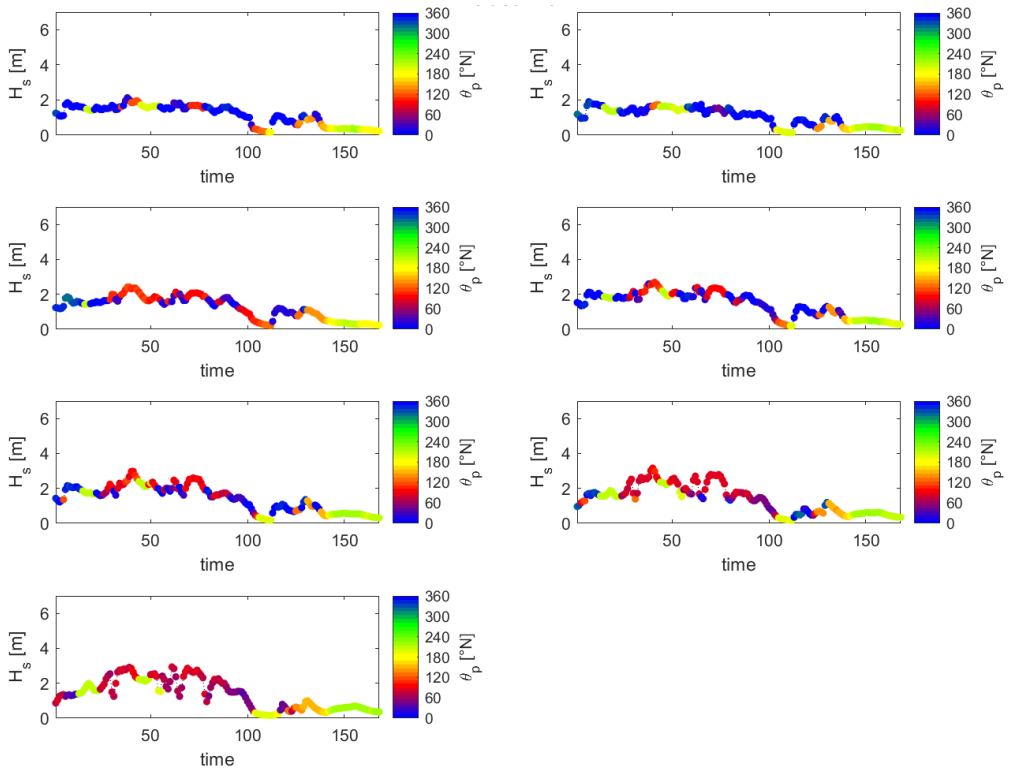

Figure 17 : État significatif : caractéristiques d'un événement mixte de la tramontane obtenues par MDA régional.

4 Discussion

Les états du modèle résultant de l'analyse du regroupement des données saisissent efficacement les principaux modes du climat local des vagues. Les figures 9 à 12 montrent quelques exemples d'événements de libeccio et de tramontane (c'est-à-dire des vagues se propageant à partir du SO et du N), qui sont les deux secteurs dominants du climat local dans la zone étudiée dans ce travail. Cette affirmation est évidente si l'on observe la rose directionnelle de la figure 17, qui montre la distribution directionnelle des H_s selon différentes classes d'intensité.

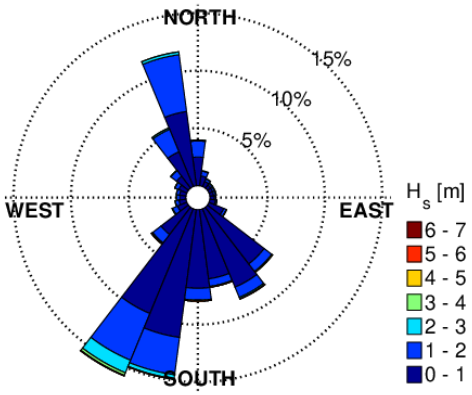

Figure 18 : Graphique polaire Hs pour le Point_000230 dans la base de données rétrospective (voir Fig. 1).

Cela se reflète dans les caractéristiques des états modèles : en fait, la plupart des états réalisés avec les deux k-means et MDA présentent des caractéristiques conformes à ces zones, tant en termes d'origine des vagues que de direction du vent. Les figures 10 et 12 suggèrent une autre considération : en observant les séries temporelles des différentes variables, on constate que l'intensité de la hauteur des vagues est proportionnelle à la vitesse du vent ; au contraire, une anti-corrélation significative caractérise les profils di $H_s - u_w$ et ceux de mslp. Cela n'est pas surprenant, puisque des corrélations similaires entre l'ensemble des variables avaient déjà été mises en évidence dans les figures 3 et 4. Veuillez noter que les séries modèles de T_p ne sont pas rapportées pour des raisons de clarté et de simplicité, mais sont fortement proportionnelles à celles de H_s.

Enfin, aucune corrélation appréciable n'a été trouvée entre les ensembles de données sur la hauteur des vagues, la vitesse du vent et la pression pour l'événement du libeccio obtenues par l'analyse des k-means (Fig. 9). Cela est probablement dû au fait que l'intensité du scénario de tempête identifié n'est pas si pertinente, même si parfois les tempêtes du libeccio devant Gênes sont caractérisées par des hauteurs de vagues allant jusqu'à 6-7 m. L'état sélectionné par

l'algorithme ne fait pas référence à des conditions extrêmes et orageuses et n'est donc pas déterminé par des vitesses de vent élevées ou des systèmes de basse pression.

De cette considération, on peut déduire qu'en réalité, grâce à l'analyse des k-Means, des conditions météorologiques plus douces sont définies, tant pour les états environnementaux très intenses que pour les états d'intensité réduite. En fait, les états obtenus par des k-Means se réfèrent aux conditions moyennes des données appartenant à un groupe particulier, tandis que les états sélectionnés par MDA s'approchent des bords des dispersions dans l'espace des variables. Pour expliquer ce qui précède, un exemple est donné dans un cas simplifié en 2D dans la Fig. 18.

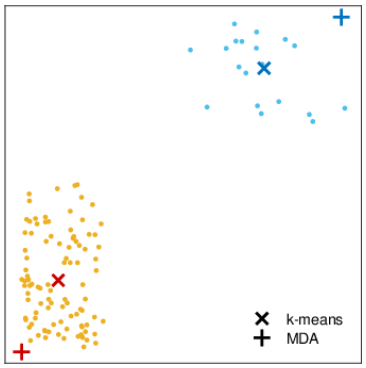

Figure 19 : Comparaison entre les centroïdes calculés avec k-Means et MDA.

Nous tenons à souligner que, lorsque l'analyse est appliquée à l'aide de l'algorithme k-Means, le profil de l'état cible de H_s ne correspond pas au centroïde calculé ; en fait, il a été choisi de sélectionner l'état le plus proche du centroïde et non le centroïde lui-même. Cela permet de considérer les états réels de la base de données rétrospective examinée comme représentatifs de chaque groupe, et non des points fictifs construits en faisant la moyenne de plusieurs variables. D'autre part, le profil de l'état cible H_s identifié avec MDA ne nécessite pas de modifications supplémentaires puisque l'algorithme ne prévoit pas de moyennage.

Ces considérations sont également valables pour l'analyse régionale. Dans les figures 13 à 16, nous observons les tendances de la hauteur des vagues en correspondance avec les événements de libeccio et de tramontane obtenus pour chaque point du sous-bassin. Les résultats permettent de voir en comparaison comment les conditions climatiques varient dans l'espace géographique avec le même état significatif considéré. En effet, il est possible de voir comment la direction de propagation des ondes varie dans les différents points analysés en fonction de leur position géographique.

Enfin, la figure 12 suggère une autre considération. On peut noter que la série de θ_p n'appartient pas seulement au secteur de la tramontane pendant toute la semaine ; en fait, il y a quelques vagues qui se transforment en se déplaçant dans le secteur du sirocco (direction entrante SE) et

du libeccio. Ce qui est mis en évidence est une caractéristique présente dans tous les états obtenus, sauf ceux liés à des événements de libeccio intense. En fait, les états météoromarins les moins aigus ne présentent pas de caractéristiques uni-modales pendant de longues périodes.

Comme on le sait, la morphologie de la côte ligurienne la rend plus exposée aux événements du libeccio (et ceci est dû au fetch de la zone examinée) : en effet, on peut remarquer que la zone examinée est caractérisée par des semaines entières déterminées par les vagues arrivant au sud-ouest. Le vent du nord et les houles de sirocco ne présentent pas des caractéristiques aussi intenses, ce qui se traduit généralement par le fait que ces événements sont généralement plus courts qu'une semaine.

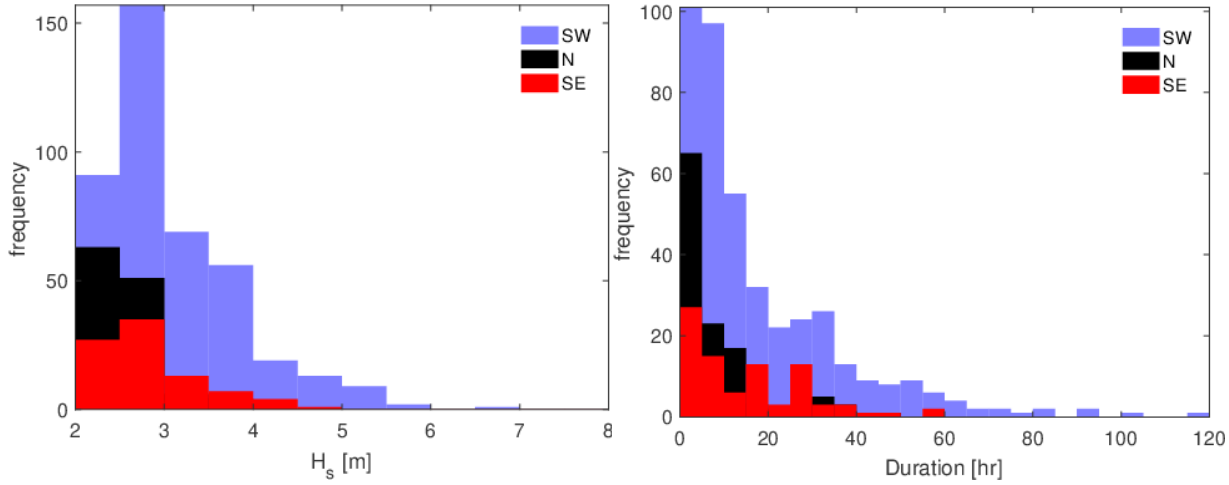

Figure 20 : Analyse de la hauteur et de la durée des ondes de tempête dans la zone étudiée.

On peut s'en rendre compte en observant la figure 19, qui montre la distribution des fréquences di H_s et la durée des tempêtes de mer effectuées au point 000230 (voir figure 1). Les tempêtes sont identifiées par une analyse des pics au-dessus du seuil (généralement appelée "POT") : c'est l'une des principales méthodes d'analyse des valeurs extrêmes et elle est basée sur l'extraction, à partir d'un enregistrement continu, des pics atteints dans une période, qui se produisent au-dessus d'un seuil. Toutes les hauteurs de vagues qui dépassent un certain seuil, fixé à 98% quantile de la distribution initiale des H_s, sont prises en considération. Par conséquent, les dépassements de H_s ont été divisés en considérant un intervalle de temps minimum entre différents groupes d'au moins 24 heures (en fait, deux groupes de hauteurs de vagues supérieures à un seuil ne sont considérés comme des événements distincts que s'ils se produisent à une distance d'au moins un jour). Enfin, une fois les événements définis, la hauteur maximale des vagues est maintenue et la durée totale de chaque groupe de dépassements est calculée. Comme le montre la figure 19, les états marins les plus intenses peuvent durer jusqu'à 5 jours pour le secteur du libeccio, tandis que dans les autres cas, ils durent au maximum

quelques jours (panneau de droite) et sont caractérisés par des valeurs H_s plus faibles (panneau de gauche).

5 Conclusions

Dans ce travail, une méthodologie a été développée pour sélectionner des scénarios climatologiques sur la base d'un ensemble de données composé de variables de la dynamique de la circulation maritime. En fait, l'objectif est de réduire le temps de calcul des modèles de dynamique des fluides, en concentrant l'attention uniquement sur certaines conditions météorologiques importantes qui sont ensuite résolues numériquement. Dans ce cas, le phénomène analysé concerne l'étude de la dispersion des microplastiques dans les eaux portuaires et les zones environnantes.

Cette méthodologie a été développée en introduisant l'application de techniques de "Data mining" par le biais d'algorithmes de clustering : en fait, le clustering est une technique d'analyse multivariée et permet de créer des groupes d'éléments, à partir de grands ensembles de données, en fonction de leur "distance logique". En particulier, k-Means et MDA ont été utilisés, qui sont deux des algorithmes les plus courants dans ce domaine.

L'application de ces algorithmes a été initialement réalisée en un seul point au large du port de Gênes et a permis d'identifier 30 scénarios climatologiques : dans le cas de k-Means, il s'agit d'états exprimant la variabilité moyenne de l'ensemble de données de départ, tandis que dans le cas de MDA, on obtient des états représentatifs également de conditions extrêmes.

Par la suite, une classification régionale a été appliquée, ce qui a permis d'identifier 30 groupes avec lesquels il est possible de décrire en détail le climat marin du sous-bassin et sa variabilité : chaque état est représentatif d'une condition climatologique particulière pour chaque point géographique de la zone.

Les résultats obtenus ont ensuite été validés en les comparant avec la climatologie moyenne du sous-bassin examiné. La validation conduit à la conclusion que la mise en œuvre de cette méthodologie peut devenir un outil utile pour la définition de scénarios multivariés permettant une réduction importante de l'effort de calcul des modèles CFD et une optimisation significative du cadre de travail en effectuant des simulations numériques plus ciblées.

Bibliografia/Bibliographie

Bárcena, J., Camus, P., García, A., & Álvarez, C. (2015). Selecting model scenarios of real hydrodynamic forcings on mesotidal and macrotidal estuaries influenced by river discharges using k-means clustering. *Environmental Modelling & Software, 68*, 70-82.

Besio, G., Briganti, R., Romano, A., Mentaschi, L., & Girolamo, P. (2017). Time clustering of wave storms in the Mediterranean Sea.

Camus, P., Mendez F. J., & Medina, R. (2011a). A hybrid efficient method to downscale wave clmate to coastal areas. *Coastal Engineering, 58*, 851-862.

Camus, P., Mendez, F., Medina, R., & Cofiño, A. (2011b). Analysis of clustering and selection algorithms for the study of multivariate wave climate. *Coastal Engineering 58 (6)*, 453-462.

De Girolamo, P., Crespi, M., Romano, A., Di Risio, M., Pasquali, D., & Sammarco, P. (2018). Wave characteristics estimation by GPS receivers installed on a sailboat travelling off-shore. *International Workshop on Metrology for the sea; Learning to Measure Sea Health Parameters (Metro Sea)*, *IEEE.*

Egbert, G., & Erofeeva, S. (2002). Efficient inverse modeling of barotropic ocean tides. *Journal of Atmospheric and Oceanic Technology, 19 (2)*, 183-204.

Enrile, F., Besio, G., Stocchino, A., & Magaldi, M. (2019). Influence of initial conditions on absolute and relative dispersion in semi-enclosed basins. *PLOS ONE, 14(7)*, 1-12.

Ferretti, G., Barani, S., Scafidi, D., Capello, M., Cutroneo, L., Vagge, G., & Besio, G. (2018). Near real-time monitoring of significant sea wave height through microseism recordings: An Application in the Ligurian Sea (Italy).

Fisher, N., & Lee, A. (1983). A correlation coefficient for circular data. *Biometrika, 70(2)*, 327-332.

Leo, F., Besio, G., Zolezzi, G., & Bezzi, M. (2019). Coastal Vulnerability assessment: through regional to local downscaling of wave characteristics along the bay of Ialzit (Albania). *Natural Hazards and Earth System Sciences 19 (1)*, 287-298.

MacQueen, J. (1967). Some methods for classifications and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability.* (p. volume 1, 281-297). Oakland, CA, USA.

Mentaschi, L., Besio, G., Cassola, F., & Mazzino, A. (2013). Developing and validating a forecast/hindcast system for the Mediterranean Sea. *Journal of Coastal Research 65 (sp2)*, 1551-1557.

Mentaschi, L., Besio, G., Cassola, F., & Mazzino, A. (2015). Performance evaluation of wavewatch iii in the mediterranean sea. *Ocean Modelling, 90*, 82-94.

Mucerino, L., Albarella, M., Carpi, L., Besio, G., Benedetti, A., Corradi, N., & Ferrari, M. (2019). Coastal exposure assessment on Bonassola bay. .

Nash, J., & Sutcliffe, J. (1970). River flow forecasting through conceptual models part i--a discussion of principles. *Journal of hydrology, 10(3)*, 282-290.

Núñez, P., García, A., Mazarrasa, I., Juanes, J. A., Abascal, A. J., Méndez, F., & Medina, R. (2019). A methodology to assess the probability of marine litter accumulation in estuaries. *Marine Pollution Bulletin, 144, 309-324*.

Re, C., Manno, G., Ciraolo, G., & Besio, G. (2019). Wave energy assessment around the aegadian islands (sicily). *Energies 12(3)*, 333.

Richter, B., Baumgartner, J., Powell, J., & Braun, D. (1996). A method for assessing hydrologic alteration within ecosystems. *Conservation biology, 10(4)*, 1163-1174.

Richter, B., Baumgartner, J., Wigington, R., & Braun, D. (1997). How much water does a river need? *Freshwater biology, 37(1)*, 231-249.

Ricther, B., Baumgartner, J., Braun, D., & Powell, J. (1998). A spatial assessment of hydrologic alteration within a river network. *Regulated Rivers: Research & Management: An International Journal Devoted to River Research & Management, 14(4)*, 329-340.

Sartini, L., Besio, G., & Cassola, F. (2017). Spatio-temporal modelling of extreme wave heights in the Mediterranean Sea. *Ocean Modelling, 117, pp. 52-69*.

Sartini, L., Besio, G., Dentale, F., & Reale, F. (2016). Wave hindcast resolution reliability for extreme analysis. *The 26th International Society of Offshore and Polar Engineering Conference*, *International Society of Offshore and Polar Engineers.*

Solari, S., & Alonso, R. (2017). A new methodology for extreme waves analysis based on weather-patterns classification methods. *Coastal Engineering Proceedings, 1(35)*, 23.

Willmott, C. J. (1981). On the validation of models. *Physical geography, 2(2), 184-194.*

Zughayar, R., Gudmestad, O. T., De Leo, F., & Besio, G. (2017). Metocean Extreme Estimations: The Sensitivity of Offshore Design Measures to Statistics' Uncertainties. *The 27th International Oean and Polar Engineering Conference*, *International Society of Offshore and Polar Engineers.*