Final Year Project Report

Revisiting the sea state bias modelling to improve radar altimeter measurements of sea level

Héloïse Lafargue

Master of Computer Science, Imaging and Multimedia, ENSEEIHT Engineering school
Apprenticeship in Data Science & AI applied to Altimetry at CLS, Toulouse
September 2023 - September 2024

Contents

Acknowledgement

I would like to express my deepest gratitude to my tutor, Ngan Tran, for her guidance, the independence she allowed me to have, her assistance in navigating various challenges, her kindness, availability, and her expertise in SSB and neural networks.

I am very thankful to my manager, Laïba Amarouche, for her unwavering support, her expertise, and for introducing me to altimetry.

I am also grateful to my team members, Estelle Mazaleyrat, Geoffroy Bracher, Giovanni D'Apice, Morgane Farradeche, Mohamed Mrad, and Thibault Pirotte, for the moments we shared and their support.

Additionally, I want to thank everyone I met during my apprenticeship, especially Marie Cherrier and the orbit determination team, Nicolas Richard for making this experience incredibly enriching in a field that was previously unfamiliar to me.

Special thanks to the CLS company and Jean-Pierre Florens and Jean-Michel Loubes, the mathematics professors who contributed to the mission with their expertise.

I am also grateful to my school, ENSEEIHT, for providing excellent education in computer science and fostering adaptability. My sincere thanks goes to my school tutor, Sylvie Chambon, for her dedication, trust, attentive support, and invaluable assistance both now and in my future professional life.

Thank you all for this new and rewarding experience and for your trust.

1 Professional Context

Company Presentation

CLS (Collecte Localisation Satellites) is a subsidiary of the French Space Agency (CNES) and CNP, specializing in satellite-based data collection and location services. Established in 1986, CLS provides innovative solutions to support sustainable management of our planet's resources. The company operates in various sectors, including environmental monitoring, maritime security, and climate studies (Fig. 1). CLS employs 900 people at its headquarters in Toulouse and at its 30 other sites worldwide.

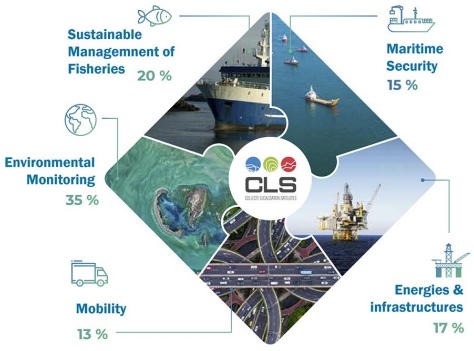


Figure 1: Sectors of activity of CLS

Work Environment

During my internship, I have been part of the Business Unit Environment & Climate, specifically within the Satellite Observation Division, Altimetric System team. The altimetry team focuses mainly on sea surface height retrieval and other associated oceanographic parameters using satellite data. These data are crucial for various applications, including climate monitoring, weather forecasting, and maritime safety.

Objectives of the Project

The primary objective of my work at CLS is to develop a new method for calculating the Sea State Bias (SSB) using five variables. SSB is a critical correction applied to altimetry measurements to account for the impact of ocean waves on the radar echo received by altimetry satellites. Accurate SSB correction is essential for improving the precision of sea level measurements and enhancing our understanding of ocean dynamics.

Project Context

Currently, existing methods for calculating SSB rely on a limited number of variables (up to 3), which may not fully capture the complexities of sea state conditions. By introducing additional variables, the goal is to refine the SSB behavior description and achieve more accurate altimetry data. This project builds on previous research and methodologies in the field of satellite altimetry, aiming to enhance the precision of oceanographic measurements and contribute to the broader efforts in climate monitoring and environmental protection.

As part of the apprenticeship, I am benefiting from a collaboration established by CLS with two mathematics professors, Jean-Pierre Florens from TSE and Jean-Michel Loubes from Paul Sabatier University. Jean-Pierre Florens is notably the mathematician who collaborated with Philippe Gaspar (at CLS twenty years ago) to develop the empirical approach that is still used today to calculate SSB solutions implemented in operational altimetry chains.

2 Introduction

The ocean, covering 71% of our planet's surface and containing 96% of all available water on Earth, plays a predominant role in maintaining the global environmental and socio-economic balance. Acting as a powerful climate regulator, the oceans absorb and redistribute solar heat, with their thermal storage capacity concentrated in the first 2.5 meters of the surface, equivalent to that of the entire atmosphere. Moreover, oceans directly impact the global economy— about 50% of the world's population lives within 100 kilometers of coastlines, relying on the resources and services offered by the ocean for their subsistence and economic development [2]. Therefore, precise ocean monitoring is essential for understanding the impacts of climate change, particularly the rising sea levels that result from thermal expansion and melting ice sheets.

Sea level, or Sea Surface Height (SSH), is a critical climate variable monitored by the Global Climate Observing System since its increase is a visible consequence of global warming. Over the past 30 years, satellite altimetry missions have provided regular, accurate measurements of SSH, starting with TOPEX/Poseidon in 1992 and continuing through the Jason series and Sentinel-6/Michael Freilich missions. These altimetry missions have been instrumental in creating long-term global mean sea level time series, offering a precise understanding of sea level trends. Despite significant technological improvements that have reduced errors in radar altimeter SSH estimates to the order of 1-2 centimeters, challenges remain in ensuring the accuracy of these measurements.

Historically, SSB correction relied on empirical models that primarily considered the effects of wind speed and wave height through two-dimensional models. Some advanced models introduced a third dimension, incorporating the mean wave period to better describe sea state conditions. However, the advent of new altimetry processing techniques, such as delay-Doppler altimetry (also known as SAR altimetry), has renewed the need for a more detailed description of SSB corrections. SAR altimetry offers significant advantages in precision and resolution compared to traditional Low Resolution Mode (LRM) altimetry, enhancing data quality for various ocean applications. However, SAR data also introduces new challenges, including sensitivity to swell period, wave direction, and orbital wave velocities, which necessitate a re-examination of SSB correction methods.

This work proposes to revisit and refine the current SSB modeling approach to account for additional parameters beyond the classical wind speed, wave height, and wave period. By incorporating variables such as Stokes drift and wave orbital velocity, we aim to better capture the complex interactions at the sea surface that contribute to SSB. The main objective is to develop tools and innovative solutions that more accurately describe SSB behavior, particularly in SAR mode, which will ultimately improve the processing of ocean altimeter data.

The importance of accurate SSB correction extends beyond SAR mode. It is also crucial for ensuring continuity and comparability between the long-term altimeter records obtained in LRM mode and the newer SAR records. This is particularly relevant as the SAR technique, first operationalized with the Cryosat-2 mission in 2010 and subsequently used in Sentinel-3A, -3B, and Sentinel-6/Michael Freilich (S6), becomes more widely adopted. In this work, we will focus specifically on data from the S6 mission, which uniquely allows for direct comparison between SAR and LRM data, facilitating a seamless transition between these two modes and enhancing the overall accuracy of sea level monitoring.

This report will present the two new approaches developped during the mission to calculate new SSB solutions, allowing for the inclusion of additional descriptive variables more easily than is possible with the current method.

Keywords: Altimetry, Sea Surface Height, Sea State Bias, PCA, Non-parametric method, Neural Network.

3 Problem formulation

3.1 Sea State Bias

3.1.1 Altimeter context

Satellite altimetry is a key technology for measuring sea surface height (SSH), providing critical data for monitoring ocean dynamics and sea level changes on a global scale. However, the accuracy of these measurements can be compromised by Sea State Bias (SSB), a significant source of error that must be addressed to ensure reliable observations.

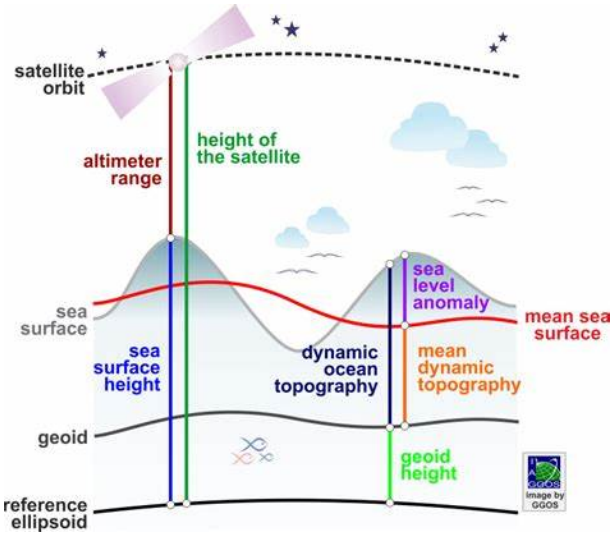


Figure 2: Definition of quantities related to sea surface heights [7]

The principle of altimetry satellite is to measure the distance (also called range) between the satellite and the surface of the Earth (Fig.2). The radar transmits a signal towards the surface that is reflected to the instrument. In the absence of waves, the reflecting surface coincides with the actual instantaneous mean sea surface averaged over the footprint. In the presence of ocean waves, the radar pulse is reflected by the inhomogeneous spatial distributions of ocean wave horizontal facets. This discrepancy between the mean sea surface and the reflecting surface modulated by waves is called the sea state bias (SSB). SSB originates from the interaction between the electromagnetic waves from the radar and the rough sea surface, leading to distortions in the altimetric readings and recorded SSH. Theoretical studies have failed to correctly describe its behavior. Part of the problem is that SSB includes also instrumental design effects as well as processing effects associated with the algorithms used to convert the return echoes into geophysical parameters (range, significant wave height and backscatter coefficient) [1].

SSB is still one of the largest sources of uncertainty in obtaining accurate estimates of SSH, leading to the biggest source of error in altimeters measurement errors if uncorrected. Fine description of the SSB behavior is crucial for improving the reliability of altimetric data, allowing a better estimation of sea level, ocean currents, and ultimately a finer grasp of the Earth's climate and its variations [10].

The measurement of SSH is not straightforward and involves several factors, it includes contributions from various components: the geoid height (h_g), which represents the Earth's gravitational field; the mean dynamic topography (η), which reflects the ocean's average surface shape due to circulation; the Sea State Bias (SSB), which introduces errors based on the sea state; and random noise (w) from the instruments [5].

The relationship between these components can be expressed by the following equation:

$$SSH = h_g + \eta + SSB + w$$

This equation highlights the necessity of correcting for SSB to achieve accurate measurements of SSH. By addressing these sources of error, the precision of altimetric data can be significantly improved. The study will focus on addressing the sea state part of the error. It leads to :

$$SSH = SSH_{uncorrected} - SSB$$

with $SSB \leq 0$ (Fig.3).

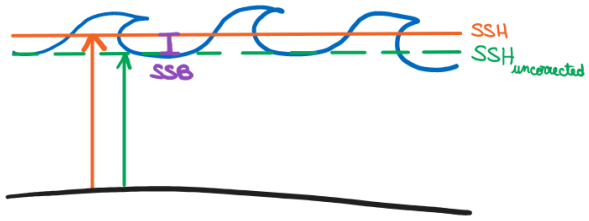


Figure 3: Schema of corrected SSH for SSB

Note that SSB is a negative value that needs to be added to the range measurement to correct it for sea state effects, since direct means to compute SSH from altimetry products is:

$$SSH = height\ of\ the\ satellite - (range + SSB + iono) - DTC - WTC - DAC - tides$$

Where:
- Iono is the Ionospheric Correction, which corrects for the reduction in phase velocity caused by free electrons within the ionosphere;
- DTC is the Dry Tropospheric Correction to compensate the delay in the radar propagation related to the dry troposphere;
- WTC is the Wet Tropospheric Correction, which compensates for the extra delay due to water vapour and liquid water in the atmospheric path;
- DAC is the Dynamic Atmospheric Correction, and includes both the static response (also known as the Inverse Barometer Effect) and changes associated with the ocean's response to sea level pressure and winds;
- tides term consists in the response of the Earth to the gravitational disruption of the moon and sun and includes each of the contribution called respectively Ocean tide, Loading tide, Earth tide and Pole tide.

3.1.2 Observability and Access to the Quantity

The main issue is that this error cannot be directly measured over the oceans. Due to the lack of in situ reference data or theoretical modeling, SSB correction relies on empirical models calibrated using the altimetric data itself. The primary challenge in empirically determining SSB lies in extracting a signal related to sea state from SSH data, which also contains oceanic signals, residual orbit errors, geophysical and environmental corrections, and instrument-related noise. As a result, we only have access to the SSH measurements:

$$\underbrace{SSH}_{observable} = h_g + \eta + \underbrace{SSB}_{non-observable} + w$$

To estimate the SSB component, the focus must be on SSH differences, either at crossover points or along collinear tracks (Fig.4). Using SSH differences rather than the SSH measurements themselves is an effective technique to eliminate the geoid signal, which is often poorly known and difficult to model precisely. By taking the difference between SSH measurements at the same geographic location but at different times (either along collinear tracks or at crossover points), the geoid's influence, which remains constant over time, cancels out. As a result, it simplifies the estimation process:

$$SSH_2 - SSH_1 = \underbrace{(h_{g_1} - h_{g_2})}_{0} + (SSB_2 - SSB_1) + \underbrace{(\eta_1 - \eta_2) + (w_2 - w_1)}_{\epsilon}$$

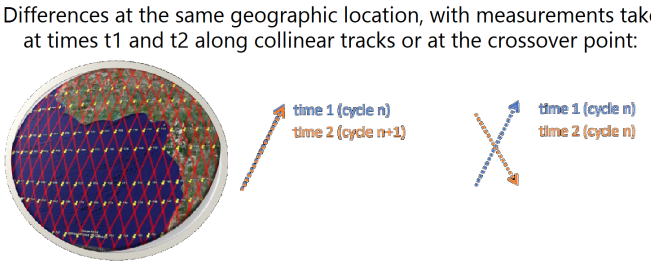


Figure 4: Collinear and crossover tracks

Indeed, the difference in mean dynamic topography ($\Delta\eta$) and random noise (Δw) over time is considered as a noise term with a zero mean (ϵ). Given that the time difference (Δt) between the two measurements is lower or equal to 10 days (depending on the dataset considered), this noise term approaches zero ($\epsilon \xrightarrow[\Delta t \simeq 10j]{} 0$).

This leads to an equation containing the term to determine and an epsilon noise component:

$$\Delta SSH = \Delta SSB + \epsilon$$

3.1.3 Mathematical formulation

The objective is to determine the functional form of the SSB model.
We introduce an arbitrary function φ, with the descriptive variables $X = (SWH, WS, MWP, ...)$, such that $SSB = \varphi(X)$.

Considering the problem with two input components, with the observations X_1 and X_2 at different moments $X_1, X_2 \in \mathbb{R}^n$, the following problem: $\Delta SSH = SSH_2 - SSH_1 = SSB_2 - SSB_1 + \epsilon$ is rewritten as a regression equation:

$$z = y_2 - y_1 = \varphi(X_2) - \varphi(X_1) + \epsilon$$

The associated minimization problem in φ is:

$$\min_{\varphi} \|z - (\varphi(X_2) - \varphi(X_1))\|^2$$

Ideally, we would have the same function, φ, regardless of the sea state measured at a given time t. However, in practice, we find that using sea state measurements at time 1 or time 2 yields different solutions. This discrepancy is likely because epsilon is not a true zero mean noise, necessitating the estimation of two functions, φ_1 and φ_2, and then averaging them (Fig.5).

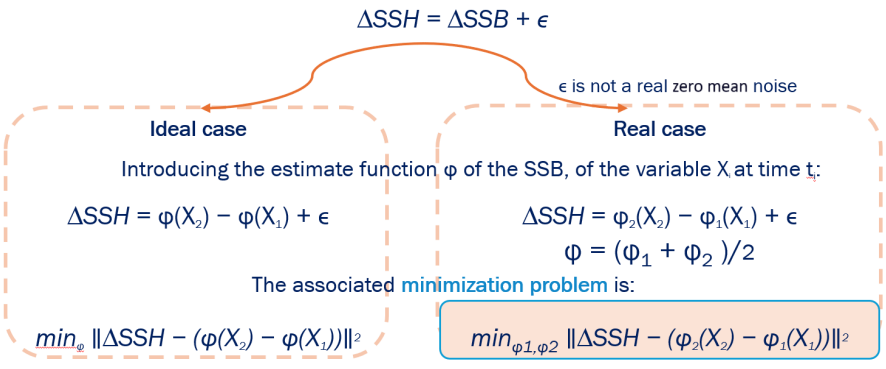


Figure 5: Introducing a function φ

This results in the following minimization problem :

$$\min_{\varphi_1, \varphi_2} \|\Delta SSH - (\varphi_2(X_2) - \varphi_1(X_1))\|^2$$

3.1.4 Selection of 5 new parameters

Traditionally, this problem has been solved using a non-parametric method based on kernel smoothing, which allows for the generation of 2D or 3D grids of SSB solutions using variables such as wave height, wind speed, and wave period. However, this solution is now reaching its limits, as there is a need for finer-scale modeling. With the advancement of altimetric instruments, we now have more precise and abundant data, which drives the need to improve our model by increasing the dimensionality of the input parameters. We aim to move from a two/three-dimensional model to a four/five-dimensional model.

Regarding the selection of parameters, the goal is to incorporate the following variables into the model for a thorough understanding and accurate correction of the sea state bias:

- Significant Wave Height (SWH): a statistical measure representing the average height of the highest third of the waves observed, providing an estimation of the wave heights commonly encountered at sea (Fig. 6).

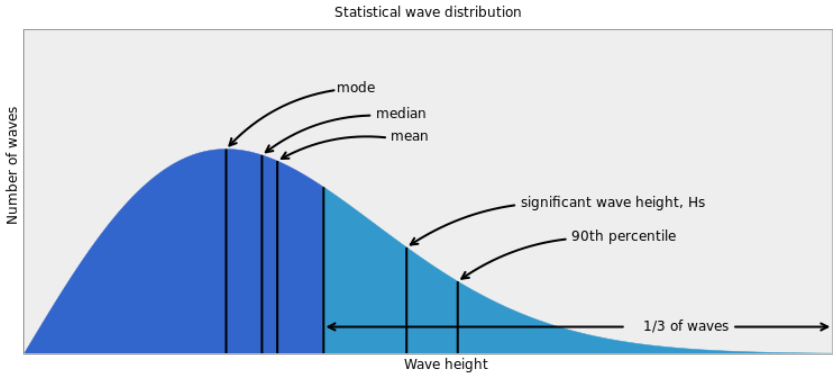


Figure 6: Statistical wave distribution and SWH [8]

- Wind Speed (WS): the speed at which air is moving in the atmosphere at a given moment, a crucial parameter that influences wave formation and, by extension, the behavior of the sea surface.

- Mean Wave Period (MWP): the average time elapsed between the passage of two successive wave crests, offering an indication of the energy and stability of wave systems at sea.

- Along Track Stokes Drifts (ATSD): measures the net movement of water particles in the direction of travel of an instrument or observer, thus capturing the impact of wave dynamics on measurements taken along a specific trajectory. It is a phenomenon of surface water transport where water particles perform orbital movements under the action of waves, resulting in a net average displacement in the direction of wave propagation (Fig. 7).

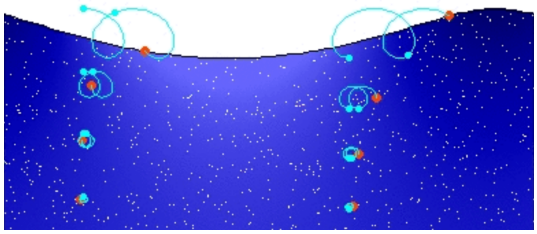


Figure 7: Stokes drift in deep water waves [9]

- Orbital Vertical Velocity (OVV): describes the speed at which water particles move in circles or ellipses under the effect of waves, a factor that influences the distribution of wave energy throughout the water column.

$$orbital\_velocity\_std = \frac{1}{4}\frac{SWH}{MWP}$$

3.2 State of the Art and Constraints

Accurate calculation of SSB is critical for reliable sea surface height (SSH) measurements, necessitating continuous refinement of current methods to meet the demands of modern oceanography.

Previously, SSB estimation often relied on parametric models, such as the BM4 model, which expressed SSB as a polynomial function of SWH and WS. The BM4 model had the form:

$SSB = SWH \times (a1 + a2 \times SWH + a3 \times WS + a4 \times WS^2)$

with coefficients $a1 = -0.04$, $a2 = 0.002$, $a3 = -0.002$, and $a4 = 8 \times 10^{-5}$ based on the empirical research on mission Jason 3. While this approach provided a foundational understanding, it was limited in capturing the complex, non-linear interactions present in real-world data.

The standard approach to SSB estimation has shifted towards a non-parametric model, developed by Gaspar and Florens (1998) [3] and later improved by Gaspar et al. (2002) [4] and Tran et al. (2010; 2021) [12, 11]. This method uses kernel smoothing to estimate SSB based on significant wave height (SWH) and wind speed (WS), allowing for greater flexibility and capturing more complex interactions compared to earlier parametric models, which were limited in their ability to represent such dynamics (Fig. 8).

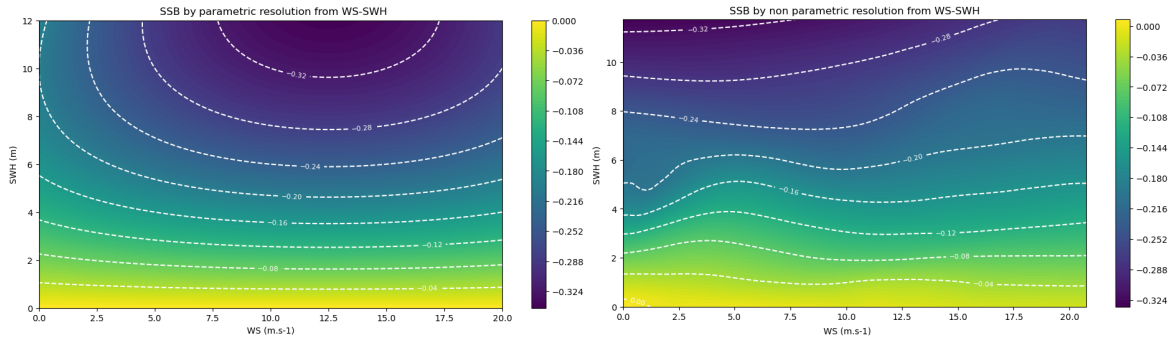


Figure 8: Evolution of the resolution method: parametric (left) to non-parametric (right)

However, the introduction of advanced altimetric instruments like SAR altimetry and SWOT radar interferometry, which offer reduced noise and higher resolution, demands further advancements. These technologies necessitate the inclusion of additional variables, leading to the development of 4D or 5D models that account for new wave dynamics, geophysical variables, and acquisition geometry.

Despite these improvements related to the use of non-parametric approach, challenges remain, including the complexity of defining smoothing parameters and the computational limits imposed by large datasets. It is currently limited to 15,000 samples/cycle for the resolution of the equation system. Additionally, ensuring smooth behaviors in SSB models where data is sparse adds another layer of difficulty. Overcoming these constraints is essential for advancing SSB modeling and enhancing the accuracy of SSH measurements in the context of modern oceanographic research.

Addressing these challenges is crucial for advancing SSB models and improving the accuracy of SSH measurements in the context of evolving oceanographic research and technology. The shift from parametric to non-parametric models marks significant progress, but further refinement and adaptation are needed to fully leverage the capabilities of modern measurement techniques.

4 Model revisit: Two new approaches

In this context, I am exploring two approaches:

- The first approach involves reducing the dimensionality of the five variables to two using principal component analysis (PCA) and then applying the traditional 2D non-parametric kernel smoothing method. This change of the input space requires adaptation of a number of parameters. This approach was thought to be quicker to implement but as the results will show below it was not that easy and also rises different questions about the interpretation of the principal component contents and how to perform the parameters adaptation.

- The second approach aims to directly uses the five variables, employing a neural network architecture with two nested models that iteratively interact to determine the two SSB functions, φ_1 and φ_2. This architecture deviates from classical models and is particularly interesting because of the minimization problem that connects the two networks.

  Note that I was only able to start deploying this approach on two variables because of lack of time to process further to five variables. The idea behind this preliminary test was to develop alternate 2D models based on the two standard variables (SWH, WS) in order to reduce the difficulty in the setting of the network and to better control the expected results. As time went fast, I was only able to deploy this new designed approach on synthetic data to verify the proper functioning of the nested networks with an iterative learning process. The definitions of an architecture, a learning dataset and the setting of the parameters for the learning step were not trivial. The different tests and associated results are reported in section 4.3. Then I will point out different aspects that might need to be consolidated before going to both real data and five input variables.

These two approaches were developed with the goal of comparing the results from the operational 2D method with the outcomes from these new methods and evaluating their potential benefits. CLS's choice was to explore these two approaches in order to identify the critical points that will need further analyses while being aware that the apprenticeship period is not long enough to fully carry out this study.

The results obtained are presented below through three sections. The first one presents results of a preliminary analysis consisting in an exploration of the different datasets along with the results of the principal component analysis performed. Then, the second section reports preliminary results on SSB modelling based on two principal components as inputs with the standard non-parametric approach (i.e. kernel smoothing). Highlights will be put on the different aspects that need to be adapted or investigated due to the input space change. Finally, the last section describes the neural network architecture proposed to estimate SSB models along with results of the first attempts based on synthetic data.

4.1 Preliminary Analysis

4.1.1 Data exploration

To conduct a comprehensive analysis, we have selected three datasets consisting in different sources for the two key parameters: SWH and WS. These latter datasets come from either the ERA-5 numerical model or from instrumental measurements taken by the Sentinel-6 satellite in either Low Resolution Mode (LRM) or High Resolution Mode (SAR). This selection allows us to evaluate whether the data source influences the analysis results, especially as we transition towards advanced altimetry technologies like SAR, for which measurement content displays some differences with the nominal LRM mode. SAR SWH estimations are for instance sensitive to wave orbital velocity while LRM data are not.

The SAR technique, first used operationally on Cryosat-2 mission, offers higher resolution and precision compared to conventional LRM. For this study, we focus on Sentinel-6 data, which uniquely provides simultaneous SAR and LRM measurements, allowing direct comparison and alignment with long-term altimetric data records. As altimetry measurement technologies evolve, refining SSB models becomes essential to fully exploit the precision offered by these instruments.

However, before directly applying new modelling approach to SAR data, it is important to validate these approaches using the well-masterized LRM data as a benchmark. This enables us to identify which differences come from the new modelling approach and which ones come from the change of data source.

All analyses are carried out over a period of one year, throughout 2022 (from cycle 42, pass number=73 to cycle 79, pass number=23), with data from Sentinel 6 JTEX mission. A filter is used to select only valid data (flag_val.alti), this flag is defined during CLS cal/val activities to identify anomalous data in order to discard them in studies.

Statistics and normalisation

The statistical analysis of the five variables reveals substantial differences in means, standard deviations, and ranges (mean $\in [0.10, 8.31]$, standard deviation $\in [0.04, 3.75]$), highlighting the need for some data preparation (Fig. 9). Without proper standardization, certain variables would disproportionately influence the analysis due to their scale. Therefore, standardization is applied to ensure balanced contributions from all variables before proceeding with further analysis.

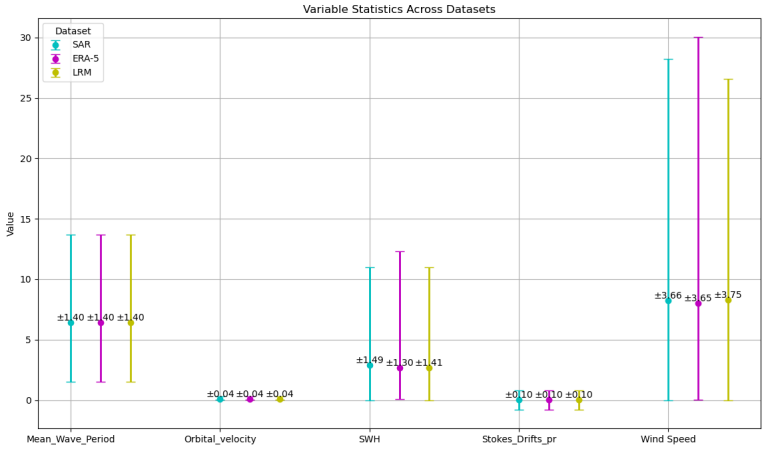
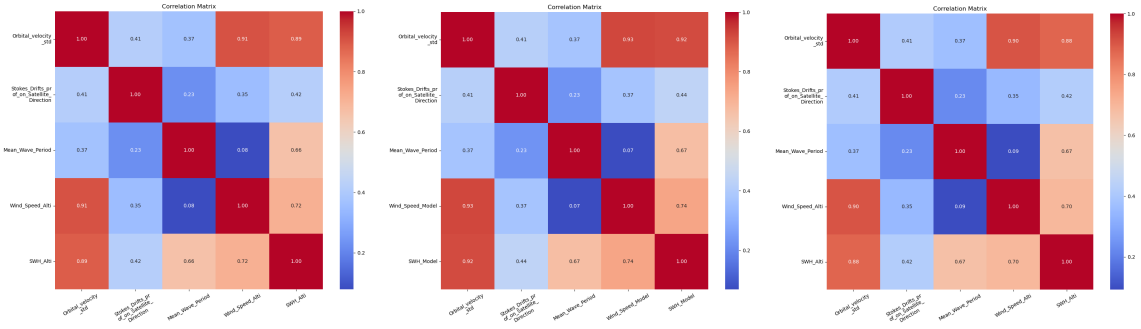


Figure 9: Statistics of the datasets (mean, max, min, standard deviation)

Correlation analysis

The correlation analysis (Fig. 10) shows strong and consistent relationships across the three datasets between orbital velocity and both wind speed and significant wave height. This indicates a robust link between these variables, independently of the data source. These insights could be usefull for understanding the influence of each variable on the SSB behavior.



(a) Set 1 (SAR) (b) Set 2 (ERA-5) (c) Set 3 (LRM)

Figure 10: Correlation matrix between the 5 variables for the three datasets

4.1.2 Principal Component Analysis

In order to address the complexity and high dimensionality of the variables involved in SSB estimation, we will apply Principal Component Analysis (PCA). PCA is a powerful statistical technique that simplifies the dataset by transforming the original variables into a set of uncorrelated components. This reduction in dimensionality will enable us to focus on the most significant variations in the data while preserving the essential information needed for further analysis.

An additional objective of this transformation is to enable the reuse of the current 2D-modelling approach. By reducing the dimensionality of the data, we can apply the existing methods to the newly derived principal components, ensuring that the refined methodology builds on established practices.

Before presenting the PCA results, we will first quickly recall the methodology behind this approach.

Principle

The goal of PCA is to find q principal components $C_1, ..., C_q$ with $q \ll p$ (where p is the number of original variables) as new variables that are linear combinations of the original variables $x_1, ..., x_p$ such that the C_k are pairwise uncorrelated, of maximum variance (to maximize dispersion to retain characteristics), and of decreasing importance.

Steps in finding the principal components:

- **Data standardization :** $X \to X_{standardized}$
  Before calculating the principal components, standardization is crucial when variables-features have different units or range. It refers to the process of scaling data so that each variable has a mean of 0 and a standard deviation of 1. This step ensures that each variable contributes equally to the analysis, preventing variables with larger scales from dominating the principal components.

- **Calculation of the covariance matrix** : $\Sigma = \frac{1}{\#data} X^T_{standardized} X_{standardized}$
  The covariance matrix plays a central role in PCA because it contains all the necessary information on the variance and correlation structure of the data. The diagonal elements of this matrix represent the variances of each variable, while the off-diagonal elements represent the covariances between pairs of variables.
  Variance measures the dispersion of data around their mean, and covariance measures how two variables vary together. PCA seeks to redirect the axes of the coordinate system to align these axes with the direction where the data extend the most, which is directly related to the variance and covariance of the data.

- **Eigenvalue decomposition of the covariance matrix :**
  Find the solutions of the characteristic polynomial $\chi_\Sigma(\lambda) = det(\Sigma - \lambda I_n)$ to obtain the eigenvalues λ_i of Σ. The λ_i are sorted in descending order of importance, reflecting the decreasing order of the variance explained by each principal component.
  The eigenvectors determine the direction of the principal components, and the eigenvalues indicate the magnitude of the variance along each component. Sorting these eigenvalues in descending order ensures that the first principal component captures the most variance.

- **Principal component selection :**
  Principal components are selected based on the eigenvalues, which reflect the importance of each component in explaining the variance in the data. The largest eigenvalues λ_i associated to the eigenvectors v_i define the directions of greatest variance, i.e. the direction in the space of variables where the data varies most. Principal components are those directions, or axes, that capture the most variance.

- **Data projection on the new components :** $C_i = X_{standardized}\ v_i$
  Finally, the standardized data are projected onto the selected eigenvectors, where i typically ranges from 1 to 2 to start the analysis. This step transforms the data into the principal component space, where each new axis (component) is uncorrelated and ordered by the amount of variance it explains.

In summary, the covariance matrix and its eigenvectors are essential in PCA because they allow for the reorganization of the data space in a way that highlights the most significant directions (in terms of variance). This enables the reduction of data dimensionality while preserving as much variance (information) as possible. This will facilitate the analysis and interpretation of complex structures in the data.

Results

First, we focus on a global analysis, meaning we examine correlations between variables across the entire globe, looking at the average behavior without geographic concerns (i.e. not looking at particular behavior).

Components	**Set1 (SAR)**	**Set2 (ERA-5)**	**Set3 (LRM)**
2	82.90%	84.01%	82.58%
3	98.03%	99.00%	97.69%

Table 1: Explained variance with N = 2 or 3 components for different datasets

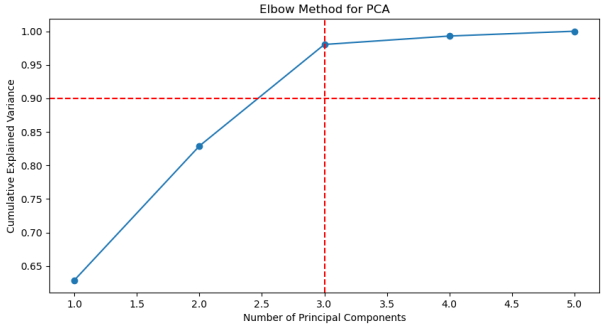


Figure 11: Elbow method on dataset SAR

For the three datasets, the analysis shows that two components explain more than 82% of the variance, and moving to three principal components would explain almost 98% (Table 1 and Fig.11 with red lines indicating cumulative explained variance exceeding 90%).

	SET 1 (SAR)		SET 2 (ERA-5)		SET 3 (LRM)	
Variable	PC1	PC2	PC1	PC2	PC1	PC2
Orbital_velocity_std	**0.954**	-0.205	**0.961**	-0.196	**0.950**	-0.214
Stokes_Drifts_prof_on_Satellite_Direction	0.567	-0.015	0.573	-0.019	0.568	-0.030
Mean_Wave_Period	0.550	**0.822**	0.542	**0.832**	0.560	**0.814**
Wind_Speed	**0.836**	-0.500	**0.845**	-0.495	**0.831**	-0.499
SWH	**0.956**	0.178	**0.968**	0.173	**0.951**	0.188

Table 2: Loading coefficients for PC1 and PC2 of the three datasets

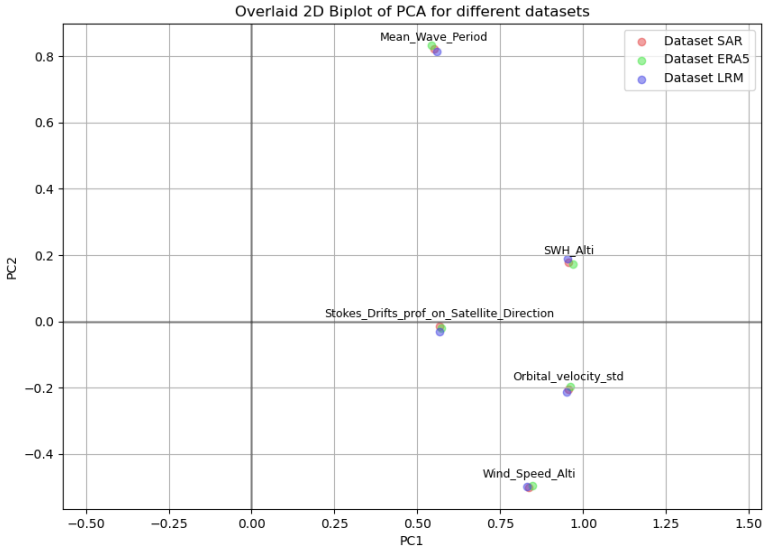


Figure 12: Comparison of PCA biplots for different datasets, on global scale

In both biplot in Fig.12 and Table.2, we can observe the contribution of each variable to the two components. It appears that the orbital velocity of waves, significant wave height, and wind speed are significant parameters since they contribute greatly to the first component. The wave period is also an important parameter, especially for the second component.

It seems that the data source does not affect the correlations between parameters.

Evaluation of Local Stability

Then, we considered the local variability of physical parameters. Indeed, we will check if there is similarity between the global results and those obtained in specific areas as presented in the mapping of Figure 13.

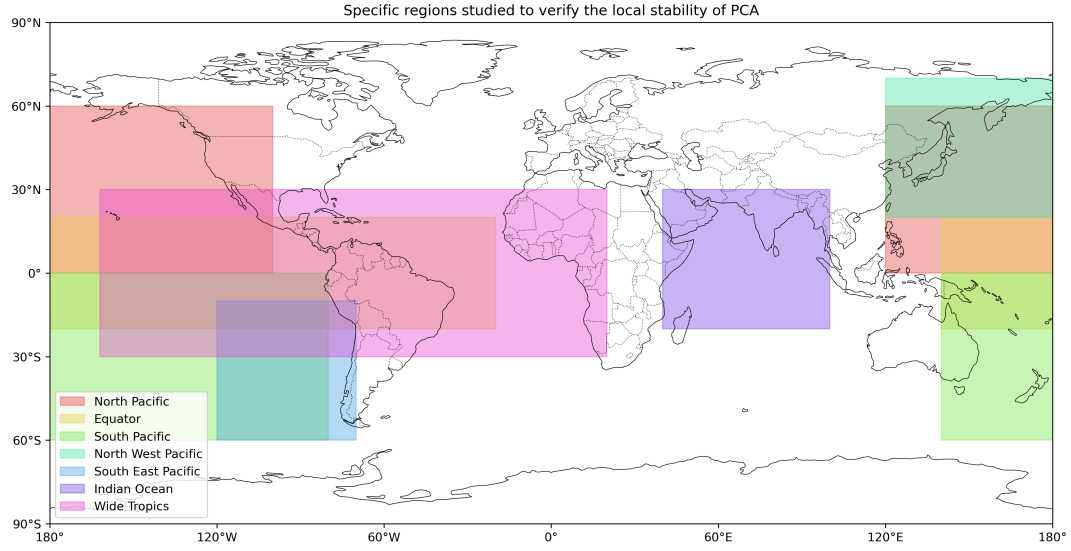
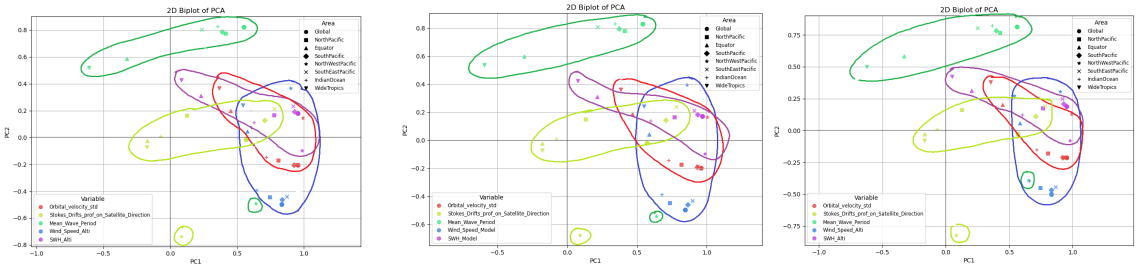


Figure 13: Mapping of specific regions studied to asses the stability of PCA results

In Fig.14, we show the biplot for each dataset with contributions to the two principal components by regions-based data, with each group of variables encircled per color to facilitate the reading of the results by visualizing the dispersion due to the regions delimitation. The details of the loading coefficients, with the predominant ones in bold, are given in appendix by Tables 4, 5, and 6.The region-based results confirm the main influence of SWH, WS, and OVV on the first

component and MWP on the second component, except for the equatorial and tropical regions which show different behaviors.



(a) Set 1 (SAR) (b) Set 2 (ERA-5) (c) Set 3 (LRM)

Figure 14: PCA biplot based on regional subsets

This analysis reveals only very small differences between the datasets, indicating that the overall patterns and behaviors are similar across different data sources. This similarity indicates that refined methodologies developed for LRM data will likely be easily transposable to SAR data.

Based on this statistical and correlation analysis, we can conclude that it seems reasonable to focus on the LRM dataset first to develop modelling strategies. Indeed, the different datasets exhibit equivalent characteristics. And it allows for direct comparison with the standard SSB solution, which was also calculated using LRM. Moreover, the results from the principal component analysis reported below will also confirm that the three datasets display very similar characterictics. This choice maintains a level of simplicity and comprehensibility before introducing additional complexity.

4.2 First proposition: Replacing (WS, SWH) with (PC1, PC2) in the standard 2D SSB model

To simplify the modeling process while incorporating the information from all five descriptive variables, we use Principal Component Analysis (PCA) to reduce the input space to two dimensions. This allows the use of the standard 2D non-parametric kernel smoothing method by incorporating the key variability displayed by the five variables, in other words enabling us to apply a familiar approach while still benefiting from a richer sea-state description.

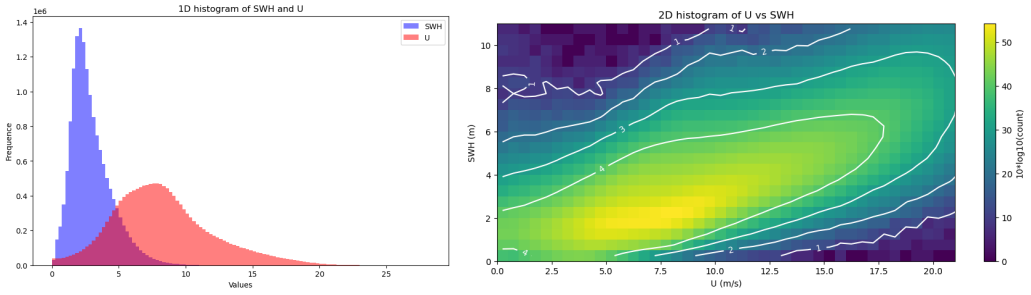
Reduced Data

Based on the PCA results, we select the two first two principal components, PC1 and PC2, which encapsulate the condensed information with more than 80% of the variance of the five original variables.

When we compare the distributions of PC1 and PC2 (Fig.16) to those of WS and SWH (Fig.15), we observe that PC1 and PC2 exhibit different density distributions. These differences point out that adaptations are required in the non-parametric method to account for the new input data structure. The statistics for each dataset are also provided in Table 3.

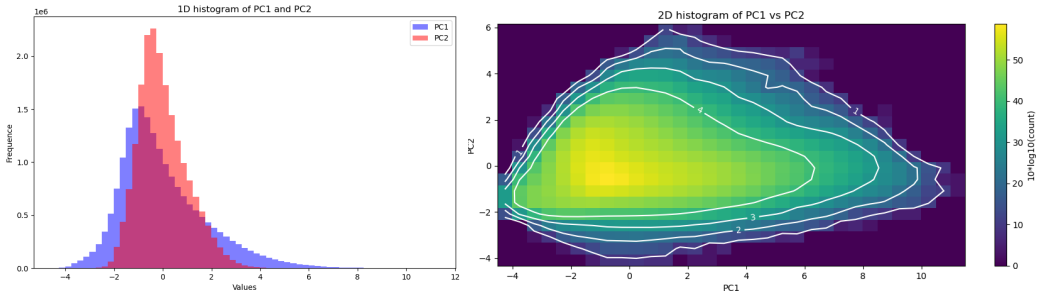
Variable	Min	Max	Mean	Std
PC1	-4.512	11.157	-7.14×10^{-18}	1.770
PC2	-4.336	5.916	-3.76×10^{-17}	0.997
SWH	0.0	11.0	2.927	1.511
U	0.0	28.22	8.27	3.673

Table 3: Statistics for Principal Components (PC1, PC2) and Standard Variables WS, SWH



(a) 1D histogram (b) 2D histogram

Figure 15: Histograms of SWH and WS



(a) 1D histogram (b) 2D histogram

Figure 16: Histograms of PC1 and PC2

4.2.1 Kernel Smoothing approach

After reducing the dimensionality of the input dataset using PCA, we apply the standard 2D non-parametric kernel smoothing method to estimate the SSB. Non-parametric kernel smoothing is a flexible technique that does not assume any specific functional form for the data, making it well-suited for complex, high-variance datasets like those involved in SSB estimation.

Recall of the motivation behind the design of this approach

Considering the problem outlined in Section 3, with two input components consisting in the observing X_1 and X_2 at different moments ($X_1, X_2 \in \mathbb{R}^n$), we have: $\Delta SSH = SSH_2 - SSH_1 = SSB_2 - SSB_1 + \epsilon$. It is rewritten as a regression equation: $z = y_2 - y_1 = \varphi(X_2) - \varphi(X_1) + \epsilon$.

The associated minimization problem in φ is: $\min_\varphi \|z - (\varphi(X_2) - \varphi(X_1))\|^2$.

Under the assumption of a zero mean noise, i.e. that $E[\epsilon|X_1, X_2] = 0$, it is equivalent to $\varphi = \arg\min_\varphi E\|z|\varphi(X_2) - \varphi(X_1)\|^2$, which is equivalent to solve (proof given in the appendix):

$$E[z|X_2 = x] = \varphi(x) - E[\varphi(X_1)|X_2 = x]$$

Introducing a regression function r, such as $r(x) = E[z|X_2 = x] = \varphi(x) - E[\varphi(X_1)|X_2 = x]$.

This can be rewritten in matrix form as a Fredholm equation of the second kind:

$$r(x) = (I_d - T)\varphi(x)$$

with $T(\varphi) = E[\varphi(X_1)|X_2 = x] = \int \varphi(x_1)p(x_1|x_2)dx_1$.

We do not know the operator T directly, but the idea is to use kernels to estimate the conditional density. Thus, this leads to an estimation of φ : $\hat{\varphi}(x) = (I_d - \hat{T})^{-1}\hat{r}(x)$. By estimating T and r, and performing matrix inversion, we obtain the estimation of φ.

Method Description

The non-parametric kernel smoothing method estimates the value of a function at a particular point by averaging the values of nearby points, weighted by their distance from the target point.

This is done using a kernel function, which assigns weights to data points based on their proximity. This is simply illustrated in Fig.17.

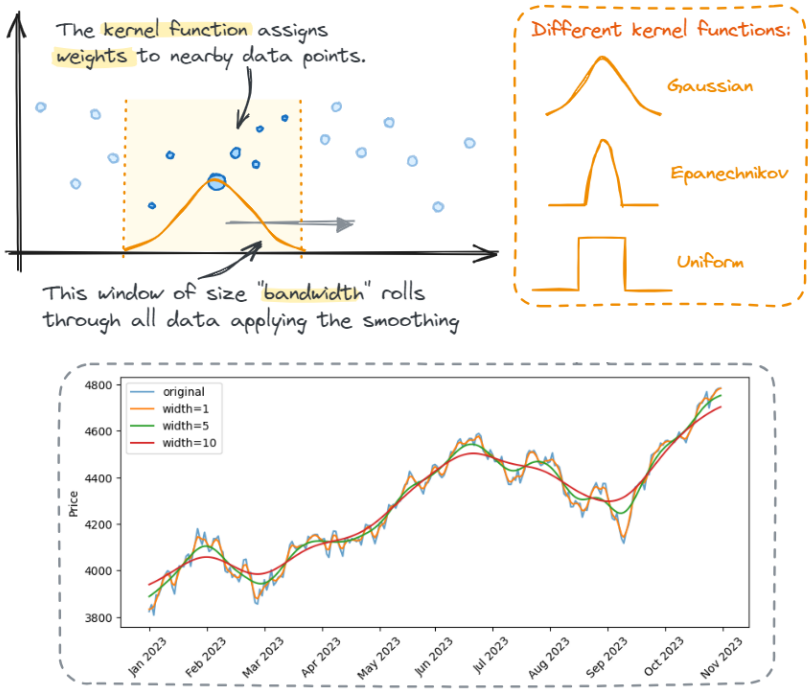


Figure 17: Kernel smoothing description on an example

In the context of SSB estimation, this method is applied to the two principal components derived from the PCA, treating them as the coordinates in the 2D space.

The key advantage of this approach is its ability to capture complex, non-linear relationships in the data without relying on predefined equations or models. This makes it ideal especially in this context where the SSB values are not directly observable and therefore difficult to model parametrically.

Key parameters in such approach

Generally, an estimator can be written as:

$$\hat{r}(x) = \frac{\sum_{i=1}^{n} K_h(x - x_i) y_i}{\sum_{i=1}^{n} K_h(x - x_i)}$$

- n is the total number of observations in the sample,
- x_i represents the values of the input variable,
- y_i represents the corresponding values of the output variable,
- $K_h(.)$ is a kernel function with a bandwidth h, which determines the weighting of observations based on their distance to x.

One of the most critical parameters in kernel smoothing is the bandwidth h of the kernel function. The bandwidth controls the size of the neighborhood around each point that is used for smoothing. A smaller bandwidth focuses on closer neighbors, resulting in a model that can capture fine details and local variability. However, if the bandwidth is too small, the model may become overly sensitive to noise, leading to overfitting. Conversely, a larger bandwidth smooths over a broader area, which can help reduce noise but may also oversimplify the model, missing important local variations.

The choice of kernel function K is another important factor. The "proximity" between points is determined by the weighting function or kernel, which assigns more weight to points near x and less weight to points farther away. Though Gaussian kernels are commonly used for their smooth,

bell-shaped curve, previous studies on SSB modelling, such as those by Gaspar et al. [4], have favored the Epanechnikov kernel:

$$K(u) = \frac{3}{4}(1 - u^2)\mathbb{1}_{\{|u| \leq 1\}}$$

The Epanechnikov kernel has a parabolic shape, which is optimal in the sense that it minimizes mean integrated squared error (MISE) for a given bandwidth. This kernel is often preferred for its efficiency, as it gives more weight to points closer to the target, which can be particularly useful in oceanographic contexts where local conditions can strongly influence measurements.

To illustrate the impact of kernel choice, we include a comparison between the Gaussian, uniform, and Epanechnikov kernels in Fig.18.

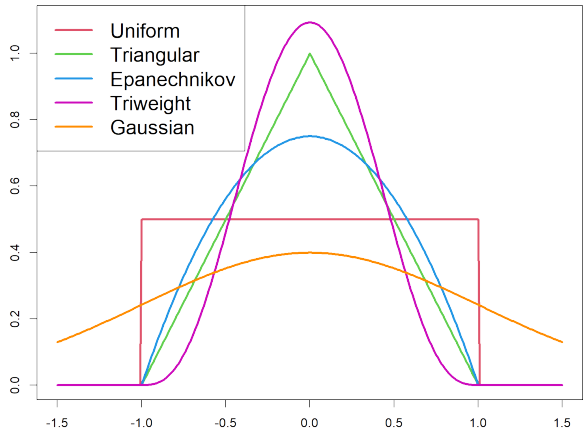


Figure 18: Comparison of kernel functions [6]

In summary, the key parameters that need to be adjusted in the kernel smoothing method include:
- **Kernel bandwidth**: Determines the level of smoothing.
- **Kernel function**: Defines the shape of the smoothing window, with the Epanechnikov kernel being chosen based on previous studies.

These parameters are optimized by testing different values to find the best-performing configuration based on the data.

Smoothing Process

The smoothing process works by assigning weights to each data point based on its distance from the point of interest, using the selected kernel function. The smoothed estimate at the target point is then calculated as the weighted average of the surrounding points. This process is repeated across the entire 2D space formed by the principal components, creating a smooth, continuous surface that represents the underlying SSB function.

This smoothing helps in making accurate predictions of SSB by effectively reducing the impact of noise and emphasizing the broader patterns and relationships in the data. It allows for a more nuanced interpretation of the SSB, capturing both local variations and overall trends, which is crucial for improving the accuracy of sea surface height measurements in satellite altimetry.

4.2.2 Evaluation criteria: Variance Difference

The evaluation of the performances of SSB solutions relies on assessing their ability to minimize the variance of SSH differences at satellite crossover points. These crossover points are locations where the satellite's ascending and descending orbits intersect, providing a natural comparison point for SSH measurements. At these intersections, the true SSH should ideally be identical, assuming accurate measurements and perfect corrections.

To evaluate the effectiveness of different SSB corrections, the SSH difference between the ascending and descending passes at each crossover point is calculated after applying the respective

SSB corrections. The variance of these SSH differences is then computed for each SSB solution. A successful SSB correction is then selected. It corresponds to the solution leading to the smaller value of the variance. This latter indicating that the corrected SSH values are more consistent between the two passes.

Mathematically, if SSH_{ref} represents the SSH corrected using a reference SSB model (e.g., $SSB_{SWH,WS}$), and SSH_{etu} represents the SSH corrected using a new SSB model under evaluation (e.g., SSB_{PC}), then the variances of the SSH differences at crossover points for these two models can be denoted as $Var(\Delta SSH_{ref})$ and $Var(\Delta SSH_{etu})$, respectively.

The idea is that if $Var(\Delta SSH_{etu}) > Var(\Delta SSH_{ref})$, then the new SSB model under evaluation SSB_{PC} is less effective than the reference model $SSB_{SWH,WS}$. This suggests that the new model does not explain as much of the variance in SSH as the reference model, making it less accurate:

$$\text{If} \quad Var(\Delta SSH_{etu}) > Var(\Delta SSH_{ref}) \quad \text{then} \quad SSB_{PC} \quad \text{is less effective than} \quad SSB_{SWH,WS}$$

Conversely, the goal is to minimize $Var(\Delta SSH_{etu})$, ideally achieving a variance lower than $Var(\Delta SSH_{ref})$. This would indicate that the new SSB model SSB_{PC} is more effective in explaining the variance in SSH, thereby improving the accuracy of the SSH measurements consists in:

$$\text{Minimizing} \quad Var(\Delta SSH_{etu}) \quad \text{to improve the effectiveness of} \quad SSB_{PC}$$

By comparing the variance of SSH differences across different SSB models, we can determine which model most effectively corrects the SSH measurements. The model that achieves the lowest variance is considered as the most accurate.

This diagnosis provides a quantitative method for evaluating and comparing SSB solutions, ensuring that the model selected for operational use is the one that best enhances the precision of SSH measurements.

4.2.3 Results and Analysis

In the current model development iteration, the dataset is segmented by observation cycles, and for each cycle, two SSB solutions —φ_1 and φ_2— are computed. These solutions are then averaged over a full year of data to produce a final SSB model that reflects the varying sea states across different regions while capturing the seasonal variability of the sea state conditions.

The model operates within a 2D framework, applying kernel smoothing techniques to interpolate SSB across a grid defined by SWH and WS. Since the model is developed within a constant because of the use of SSH differences, the final solution is shifted to fulfill the condition SSB(0,0)=0m to ensure that all SSB values are negative, aligning with physical expectations.

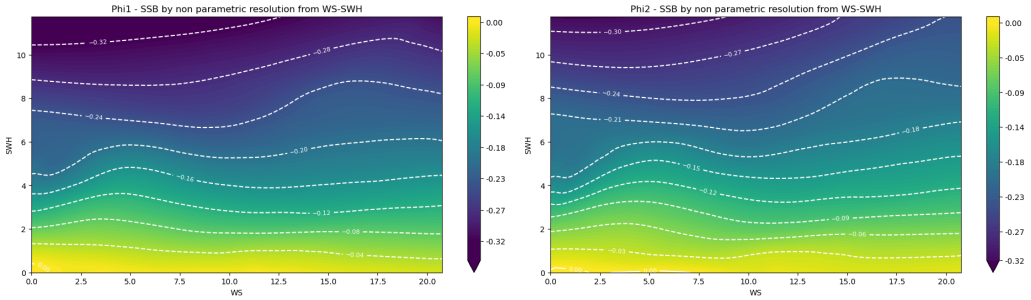
The output of the model is visualized as 2D-maps, describing the SSB behavior in this space. These maps provide valuable insights into the relationships between sea state parameters and SSB, derived from the extensive, year-long dataset.

This entire process was carried out using a Fortran version of the modeling software, which is currently being transitioned from Fortran to Python.

SSB model based on (WS, SWH)

As a baseline, we first computed a model based on WS and SWH from the S6 mission. This SSB model is homogeneous with version used in operational chains which currently provide geophysical products to altimetry users. The resulting maps are shown in Figs.19 and 20.

When one compares Figs.19a and Fig.19b, we observe that the two separate solutions, φ_1 and φ_2, exhibit slight differences. This observation underlines the importance of averaging these solutions to obtain a more stable and reliable final SSB model.



(a) φ_1 solution (b) φ_2 solution

Figure 19: φ_1 and φ_2 solutions, based on non-parametric model with variables (WS, SWH)

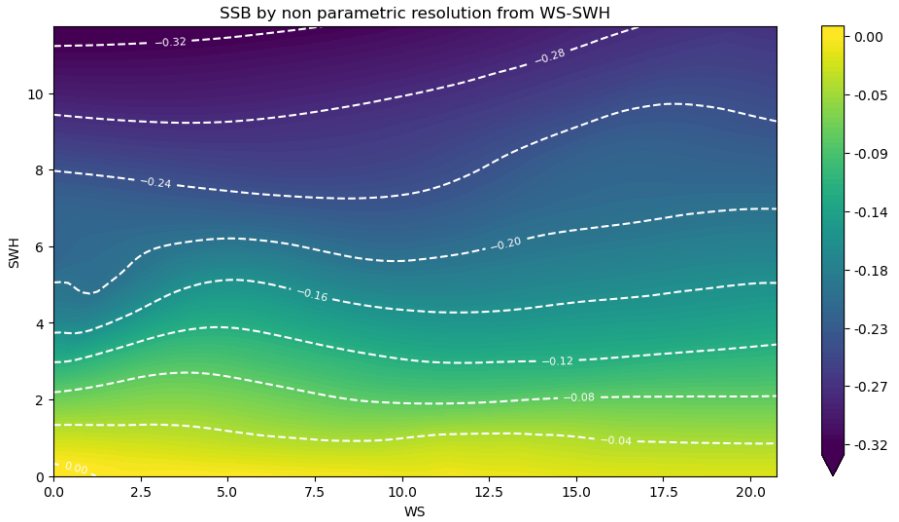
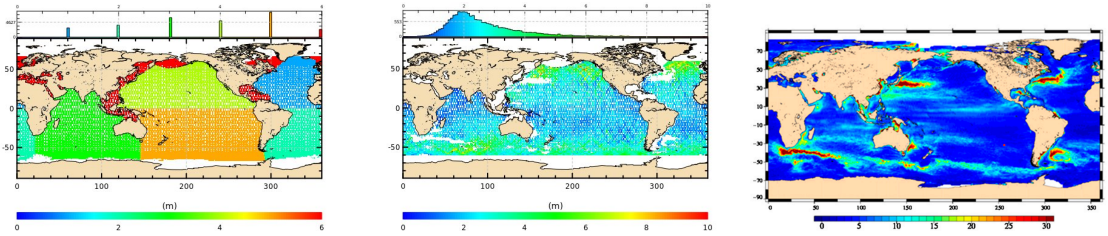


Figure 20: Final SSB version based on non-parametric model with variables (WS, SWH)

SSB models based on (PC1, PC2)

The next step consists in the computation of a new SSB model version but this time the two input data are changed. These latter corresponds to the two principal components derived from the PCA of the five variables related to SSB. By condensing the original five variables into two principal components, this approach aims to facilitate the modeling of SSB but also seek to better describe its behavior.

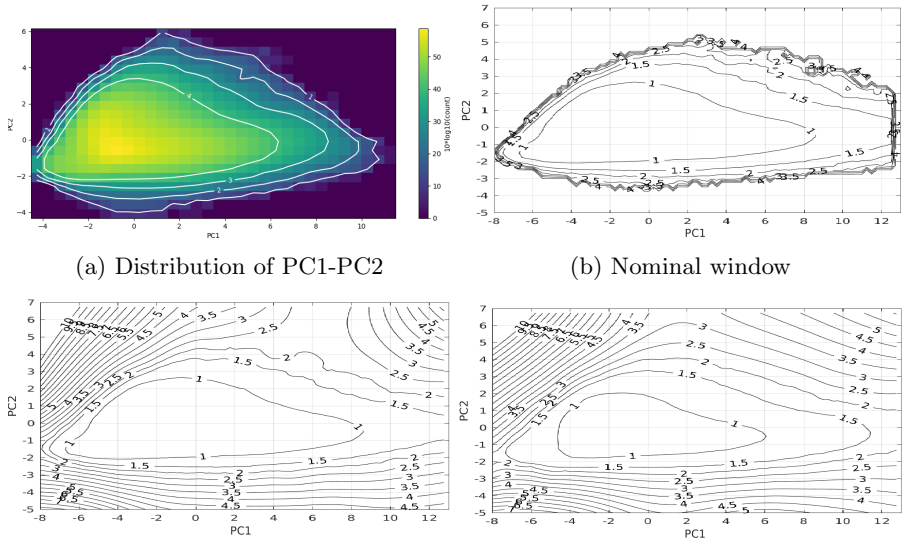
Along with the use of the cal/val flag to discard anomalous data in the study, data from specific regions were also excluded from the dataset, including enclosed seas, coastal areas, and regions marked in red in Figure 21a. Observations from the white zones in Figure 21b, which correspond to areas of very strong currents as pictured in Fig.21c, were also removed. This step is crucial to avoid including data impacted by large local effets such as interactions between waves and currents or between waves and low bathymetry such as in coastal areas that can change the wave profiles. Their impact on SSB behavior is poorly understood and are presently considered as outliers difficult to handle, therefore they are not used to focus on more simple and general sea-state conditions.



(a) Bassin number (b) Mean of SWH (c) mean currents

Figure 21: Filtered regions for the study

The computation of SSB models based on (PC1, PC2) was tested with **three different kernel windows** based on principal components distribution: one nominal and two smoother. They are presented in Fig.22 in panels b to d respectively. To find a good SSB model, we tested also **different smoothing factors values** along the two axes $h = (h_1, h_2)$, with h_1 and h_2 ranging from 1 to 3 in increments of 0.2. This resulted in a total of 121 tests, allowing us to explore a wide range of smoothing effects. The comparison of a few solutions shows that the performance of the model can be enhanced by adjusting both the kernel file definition and the smoothing parameters, by increasing the h values.

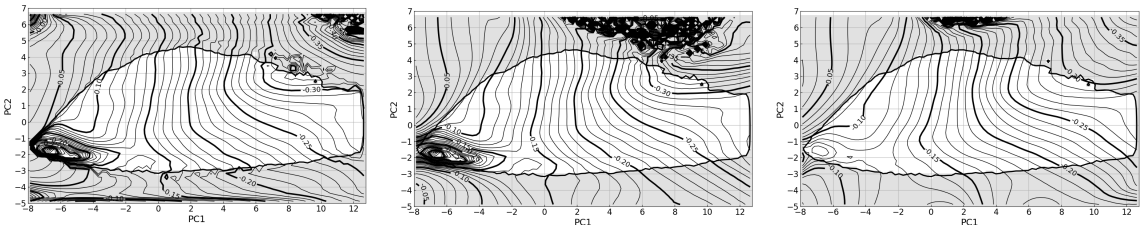


(a) Distribution of PC1-PC2 (b) Nominal window

(c) Smoothed window (d) Smoothed window bis

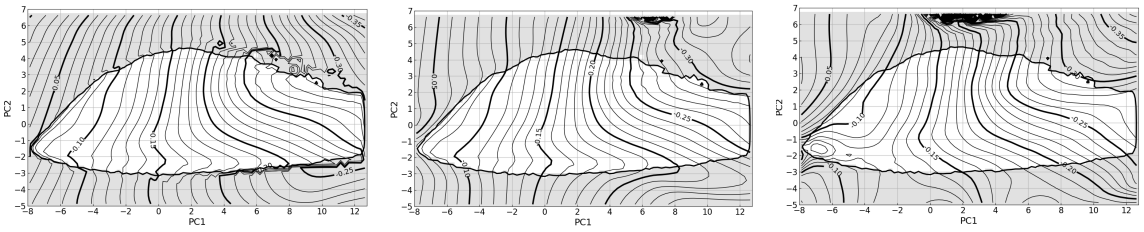
Figure 22: Kernel size grids computed from LRM data distribution

For instance, varying the kernel parameters demonstrates the challenge in achieving the best smoothing:



(a) Nominal window (b) Smoothed window (c) Smoothed window bis

Figure 23: SBB models obtained by combining $h = (2, 0.9)$ and each of the kernel size grids tested



(a) Nominal window (b) Smoothed window (c) Smoothed window bis

Figure 24: SBB models obtained by combining $h = (3, 1.5)$ and each of the kernel size grids tested

We didn't have time to evaluate / compare the performances of all computed SSB solutions. Amongst the ones we looked into, the best result in term of smoothing seems to be the one developed with the smoothed window and the bandwidht $h = (3, 1.5)$. This configuration minimized the occurrence of oscillations or ripples (Figs. 23a, 23b, 23c and 24c) and also reduced abrupt changes around the boundaries where the data population drastically declines (Fig. 23a, 24a) without smoothing to much (Fig. 24b). These abrupt changes can often indicate overfitting or inadequate smoothing, where the model becomes too sensitive to variations in the input data. By playing with these two parameters, we were able to create a more realistic and stable representation of the SSB across the grid.

Note that we need to impose a constraint to fully determine the SSB solution since the solution is computed within a constant because of the use of differences data. In the (WS, SWH) space, the natural constraint is to impose SSB = 0 when (WS=0, SWH=0), i.e. when the sea surface is flat. In the (PC1, PC2) space, it seems illogical to impose SSB = 0 for (PC1 = 0, PC2 =0) since this leads to get positive values for SSB and this is impossible by definition. This aspect will need further investigation to understand the content and meaningful of these principal components but for the time being, the decision was made to shift the global SSB values population in order to get only negative values with a distribution tail dying near 0 as shown in Fig. 27a.

Once the SSB_{PCA} solution obtained, it needs the SSB grid of values is transposed from the space (PC1, PC2) to (WS, SWH) by setting mean values for the three other variables. This allows a more direct comparison of the behaviors of SSB from these two models in the (WS, SWH) space. The principle (Fig. 25) is to start with a grid of values for WS and SWH. For each point on this grid, the principal components (PC1 and PC2) are calcultated, fixing the three other descriptive variables at their mean value. These principal components are then used to interpolate the target variable (SSB_{PCA}) based on the results obtained for PC1 and PC2.

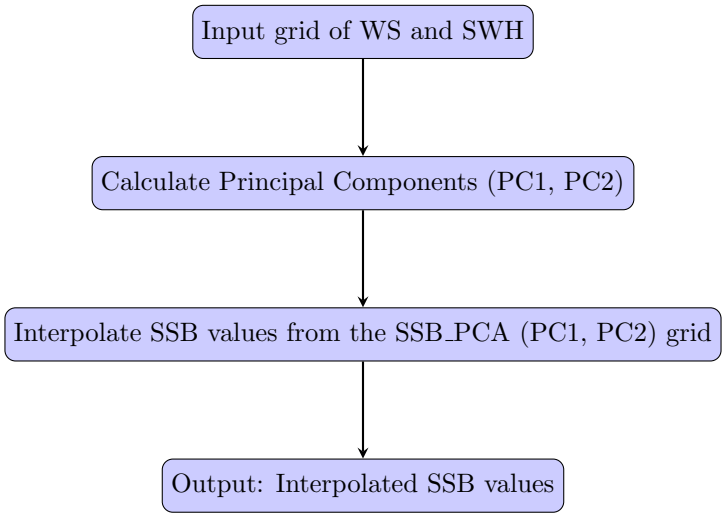


Figure 25: Steps to obtain the $SSB_{PCA}(WS, SWH)$ grid

After all of this process, we obtained the grid provided in Fig.26 and comes from the SSB_{PC} solution, with the "smoothed" kernel window and the bandwidth $h = (3, 1.5)$.

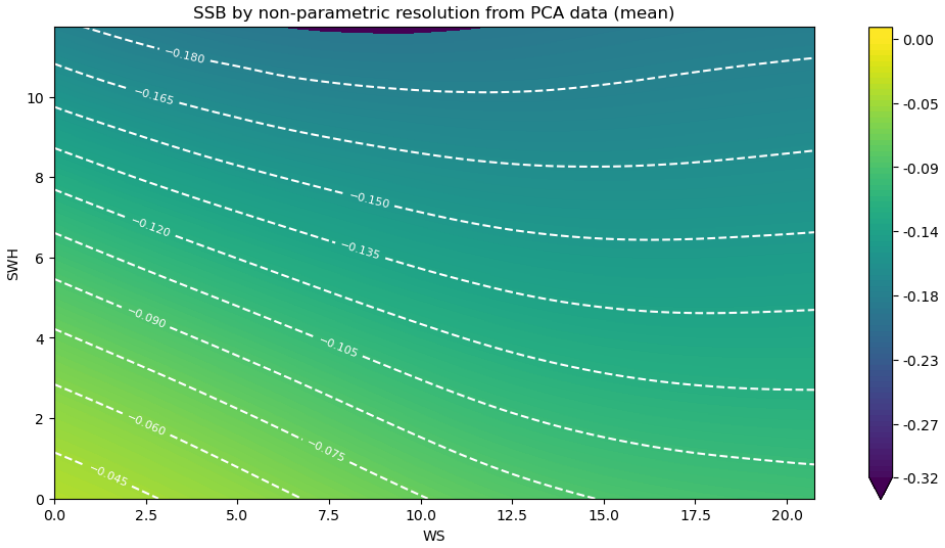
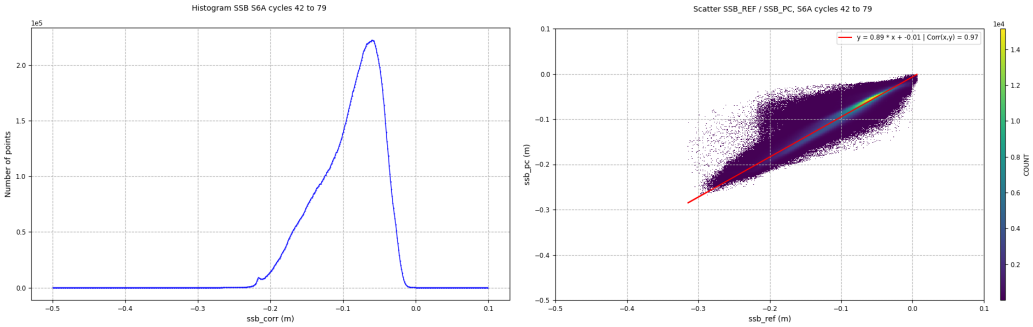


Figure 26: Selected SSB_PCA model expressed in the (WS, SWH) space

Comparison results

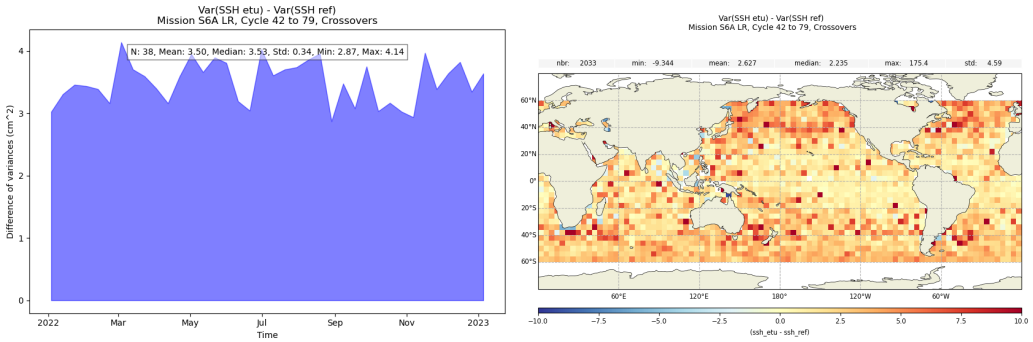
Hereafter, we compared the two SSB models developed from S6 LRM dataset: the SSB 2D calculated with (PC1,PC2), versus the SSB 2D calculated with (WS, SWH).

Comparison of the grid displayed in Fig. 26 with the one in Fig. 20 clearly show significant differences of the SSB behavior in the common (WS, SWH) space. Since SSB values are not directly observable, it is difficult at this stage to say which one is the more accurate to correct SSH data. Scatterplot provides in Fig. 27b confirms that the global features between the two models are similar since the regression fit displayed is close to the 1:1 reference line. But the data dispersion indicates also some differences in SSB estimations. Moreover, performance assessment results, provided in Figure 28, indicates that the selected SSB_PCA solution is less accurate than the reference one based on (WS, SWH) to correct efficiently the SSH data observed at crossovers.



(a) 1D histogram of SSB_{PC} solution (b) 2D histogram of SSB_{PC} vs $SSB_{SWH,WS}$

Figure 27: 1D histogram of SSB_{PC} and scatter plot of SSB_{PC} vs $SSB_{SWH,WS}$



(a) 1D histogram of SSB_{PC} solution (b) 2D histogram of SSB_{PC} vs $SSB_{SWH,WS}$

Figure 28: 1D histogram of SSB_{PC} and scatter plot of SSB_{PC} vs $SSB_{SWH,WS}$

To summarize, the observed differences in SSB behavior (Fig. 27b) and the comparatively poorer performance of the model derived from (PC1, PC2) and selected above Fig. 28a suggest that, at this stage, incorporating new variables into the SSB modeling has not yielded conclusive improvements.

Several factors might contribute to these outcomes:

1. **Parameter Optimization**: The kernel grids and smoothing parameters may not yet be fully optimized. Only a few solutions have been analyzed so far amongst the hundred versions computed. But all the components for the comparison between SSB solutions and performance results have been set up to ease the follow-up to the study. The choice of bandwidths and kernel functions are crucial in ensuring that the model accurately captures the underlying relationships in the data. The current settings might not be the best suited for this new dimensionality reduction approach.

2. **Component Selection**: By selecting only the first two principal components, which account for approximately 83% of the total variance, the model may be missing out on critical information contained within the remaining 17% of the data. This loss of information could contribute to the observed decrease in model performance. One possible way to mitigate this issue is to explore the use of Partial Least Squares (PLS) regression as an alternative or complementary technique. In PLS regression, the components are constructed specifically to best explain the response variable (in this case, the SSH), whereas in Principal Component Regression, the components are initially created solely based on the predictors variables X. This means that PLS could provide a more targeted dimensionality reduction that better captures the underlying relationship between the predictors and the SSH, potentially leading to improved the model SSB performance.

3. **Regional Variability in Correlations**: The differing correlations between the variables across various ocean basins might not be fully captured by the linear combinations used to form the principal components. In complex and heterogeneous environments like the global oceans, the relationships between variables can vary significantly by region, and this complexity might be challenging to encapsulate in a simplified PCA model.

These findings indicate that while dimensionality reduction through PCA is a promising approach, allowing for results that are close to those derived from traditional methods, further analysis and fine-tuning of the approach are necessary. Continued work is needed to optimize the parameters and explore whether incorporating additional principal components or adjusting the kernel smoothing approach can yield improved results.

4.2.4 Conclusion and Future Work

This study investigated enhancing Sea State Bias (SSB) modeling in satellite altimetry through dimensionality reduction using Principal Component Analysis (PCA) and 2D non-parametric kernel smoothing method. By reducing the five descriptive variables to two principal components, we aimed to simplify the model while retaining essential data variability.

Key Findings

- PCA Effectiveness: Although PCA reduced the dataset's dimensionality, the resulting SSB model did not yet outperform the traditional SWH and WS-based model. This suggests that crucial information may have been lost or that the linear combinations in PCA may not fully capture complex regional interactions.
- Kernel Smoothing: The choice of kernel smoothing parameters, particularly the bandwidth, significantly influenced model performance. The best results were achieved with a smoothed window and bandwidth $h = (3, 1.5)$, which minimized artifacts in the SSB behavior structure.

Future Work

- Optimizing Kernel Parameters: Future research should explore automatic bandwidth selection using techniques like Python's KeOps package, which optimizes h through gradient descent.
- Incorporating Additional Components: Adding more principal components or using Partial Least Squares (PLS) regression could capture more data variance and improve model accuracy.
- Regional Modeling: Developing region-specific models could address the varying relationships between variables across different ocean regions.

In conclusion, concerning this first attempt to improve the SSB behavior by using more than 3 descriptors, while the integration of PCA with kernel smoothing has shown promise, particularly in simplifying the model and reducing computation time, further optimization and exploration of alternative approaches are necessary to fully asses its potential in improving SSB modeling.

4.3 Second proposition: SSB estimation based on a neural network approach

The goal of this section is to present a first attempt to use a neural network-based approach to estimate SSB, leveraging the flexibility and power of deep learning models. Traditional non-parametric methods, though effective, can struggle with the complex interactions between multiple variables, especially as the dimensionality of the data increases. By applying neural networks, we expect to capture these complexities more effectively and improve by this way the accuracy of SSB estimations.

The idea behind this preliminary test was to develop alternate 2D models based on the two standard variables (WS, SWH) in order to reduce the difficulty in the setting of the network and to better control the expected results. First, this new designed approach was tested on synthetic data to verify the proper functioning of the proposed network with an iterative learning process. The section below presents the architecture definition, the learning dataset and the setting of the parameters for the learning step which were not trivial to define/choose. The different tests and associated results are also reported hereafter. Test on real data has been postponed because of lack of time during this apprenticeship. But the preliminary results are encouraging and represent definitely a good step forward.

4.3.1 Methodology with 2 variables

The approach involves designing a non-classical neural network architecture with two nested models that iteratively interact to predict SSB. This is particularly interesting because, unlike conventional neural network models that are trained end-to-end in a single optimization process, this architecture alternates between optimizing two separate models.

The core idea is to alternate between optimizing the parameters of the first model (θ_1) and the second model (θ_2), refining the predictions at each step. The loss function used in training these models is based on the difference between the observed sea surface height differences (ΔSSH) and the predicted differences from the neural network models. Through iterative updates, the models gradually minimize the error between the predicted and actual values (Fig. 29 and 30 for a more detailed view):

$$\min_{\varphi_1,\varphi_2} \|\Delta SSH - (\varphi_2(X_2) - \varphi_1(X_1))\|^2$$

This non-traditional architecture allows for more dynamic interaction between the models, where each model's predictions are used as inputs for the other, fostering a collaborative learning process. This iterative refinement makes the architecture well-suited to capture the complex, multi-variable relationships present in SSB estimation, which are often challenging for standard neural network approaches.

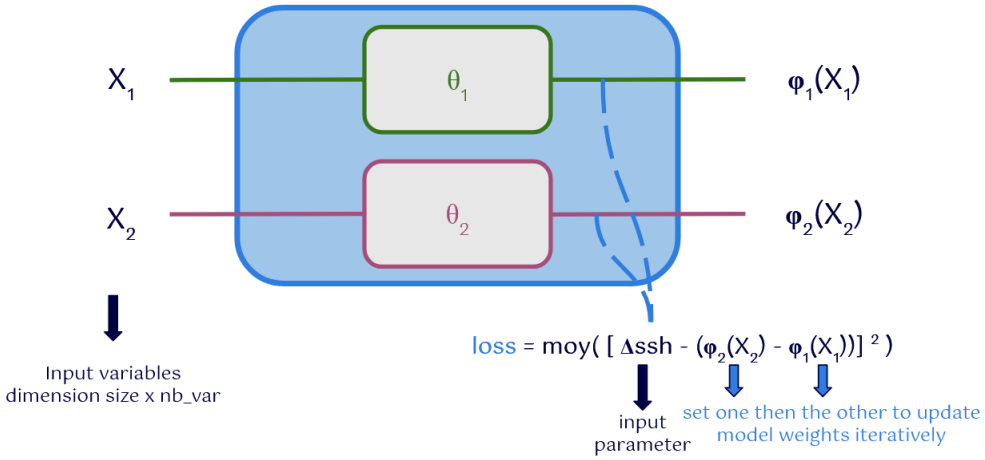


Figure 29: Simplified view of the nested neural network

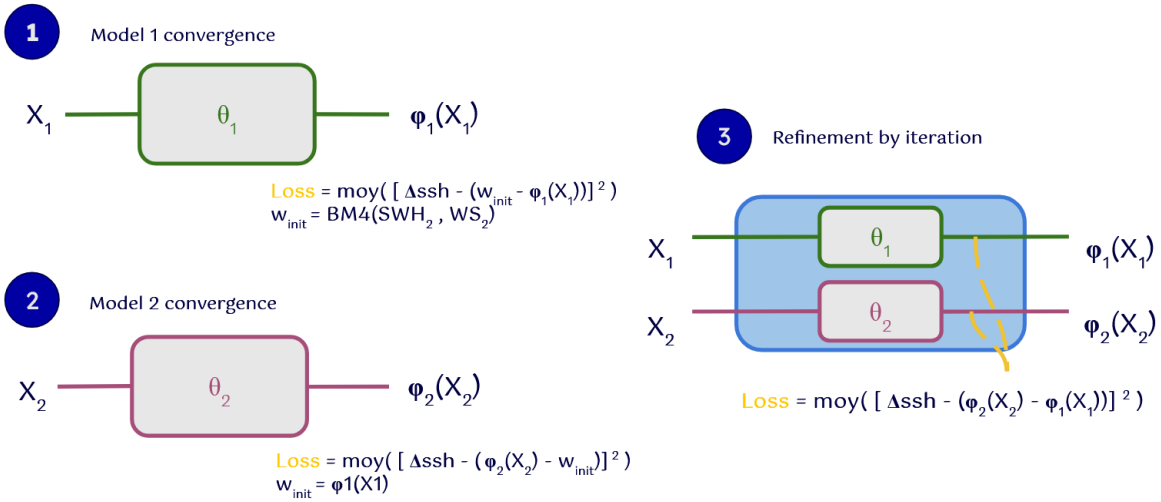


Figure 30: Full view of the nested neural network

The two input variables used in the neural network models are derived from real data, and they are normalized before being fed into the models. The dataset was not split into training, validation, and testing sets (60%, 20%, and 20%, respectively) as it is usually done to to ensure robust model evaluation. This standard practice in machine learning helps prevent overfitting and ensures that the model generalizes well to unseen data. In this early stage of the approach development, we focused mainly on how to build a good training set. The model is trained and tested on the same dataset of 15,000 samples to check its proper functioning. Of course, we will use later on the three types of datasets to validate our approach and results when we will have a better handling on the neural network methodology.

The training dataset was carefully built. We tried to have homogeneous data density over the X2 space, ensuring a balanced representation in the 2D space. However, this method of selection results in a concentrated distribution, with a peak in the X_1 space. Starting with a larger dataset of 56,744 samples, this down-sampling to 15,000 samples ensures that the data used for training is representative, while also reducing the complexity of initial tests to make the training process more manageable.

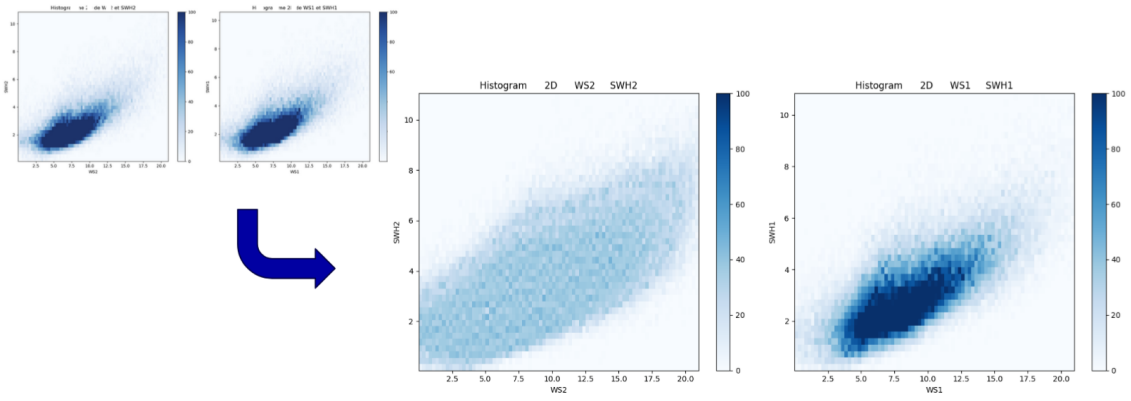


Figure 31: Data preparation: balanced representation by homogeneous distribution on X_2 space

4.3.2 Neural Network Architecture

The neural network architecture consists of several layers:

- **Input Layer**: Receives the normalized input variables.
- **Hidden Layer**: Composed of dense layers with Tanh activation functions to capture non-linear relationships in the data.
- **Output Layer**: A layer with a single neuron to provide the final SSB estimate.

We tested different configurations of hidden layers, neurons, and learning rates, systematically optimizing hyperparameters such as learning rate, and the number of hidden neurons to improve model performance. The iterative training process alternates between the training of each of the two models, updating the weights of one model based on the output of the model which has its weights frozen when the other model is trained.

4.3.3 Training and Evaluation

The training process involves initializing the weights of the two models, with several strategies tested, including random initialization, uniform initialization among others. One of the weight initialization methods used is the Xavier initialization. This method initializes the weights using a uniform distribution, optimized for networks utilizing sigmoid or tanh activation functions. By maintaining the variance of gradients across layers, it enables more stable and faster training. The models are then trained for a specified number of epochs (50 or 150) or a specified convergence

criteria ($loss \leq 10e^{-6} \Leftrightarrow SSB \leq 1cm$), and the loss function is monitored to ensure convergence.

In this initial phase, we proceed by trial and error and seek for a model based on the two standard input variablesto assess whether this approach can achieve results that are at least as good as those obtained using traditional 2D non-parametric methods. The training is conducted on synthetic data where the sea surface height difference, ΔSSH, is generated using the BM4 2D-parametric model of SSB (i.e. $\Delta SSH = BM4(X_2) - BM4(X_1)$. With

$$BM4(u, v) = v * (a1 + a2 * v2 + a3 * u1 + a4 * v12)$$

having $a1 = -0.04;\ a2 = 0.002;\ a3 = -0.002;\ a4 = 8^{-5}$.

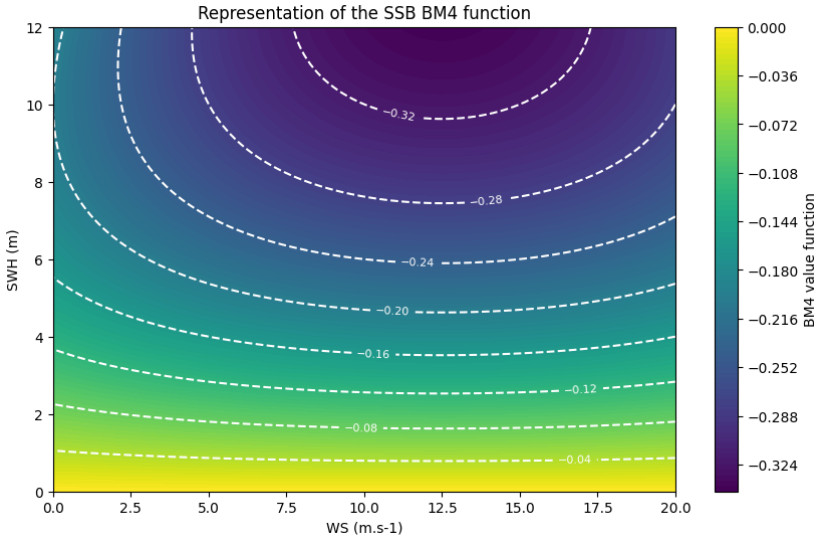


Figure 32: Parametric function of SSB BM4

This simple case helps us to determine the appropriate setting of a few parameters in this method under ideal conditions. Once the models will perform well on synthetic data, the idea is then to work with real-world data, where the challenge are significant due to the presence of substantial noise in the observed ΔSSH. This noise will complicate the training process, as the models must learn to distinguish between the true SSB signal and the underlying noise in the measurements.

Once trained, the models will then be evaluated on a test set. The key metric used for such evaluation is accuracy, measured as the percentage of predictions within a defined tolerance range of the true SSB values in this case where the expected SSB values are known, and the 2D-function shape in comparison with the known function of BM4. Several tests were conducted, varying the network configurations and hyperparameters to identify the most effective setup.

To illustrate the results of our neural network-based approach, we will present the figures as follows: The first column will display the predicted values of φ_1 and φ_2 generated by the neural network. These predictions are visualized with a fixed color scale ranging from -0.35 to 0m, which reflects the typical range of variation of the SSB values across different sea states. The second column will show the differences between the predicted φ_i values and the theoretical BM4 model, with the color scale set between -0.05 and 0.05 m. This scale is chosen to focus on deviations within a centimeter range. When the approach will be better defined, we will focus on millimeter range, which is the precision level we aim to achieve with the final model. By setting these scales, we ensure that the visual comparisons highlight the key aspects of the model's performance and its accuracy relative to the established BM4 benchmark.

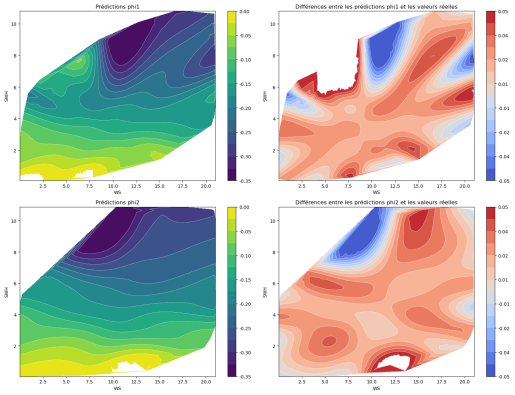
4.3.4 Results and Analysis

After implementing the neural network model with the architecture and training procedures described, the results were carefully screened to determine the effectiveness of this approach in estimating SSB.

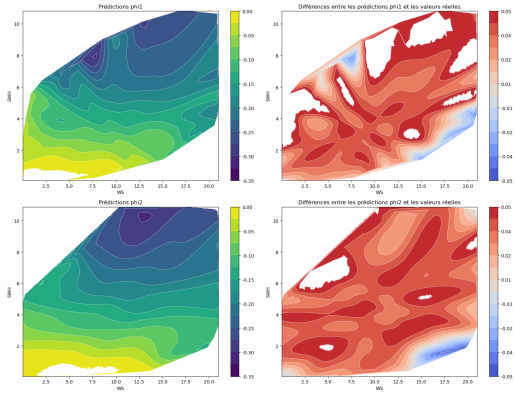
1. Performance on Synthetic Data

The initial phase of testing focused on synthetic data generated using the 2D BM4 model. This phase was essential for validating the model's ability to learn the underlying relationship between the input variables and the SSB estimations under controlled conditions.

The first test with synthetic data was conducted by training the two models separately without linking them through a shared loss function. The idea here was to quickly assess whether a simpler (Fig.33a) or more complex (Fig.33b) network could capture the underlying model, without any optimization. The results showed a slight improvement in performance when using a neural network architecture with fewer neurons in the hidden layer, suggesting that a simpler model may be more effective under these conditions.



(a) Hidden layers: Linear(16)

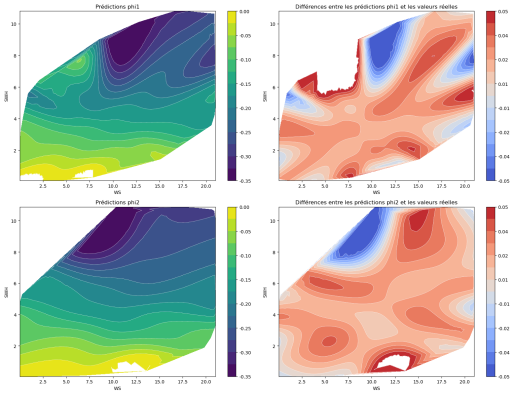


(b) Hidden layers: Linear(64) - Linear(32)

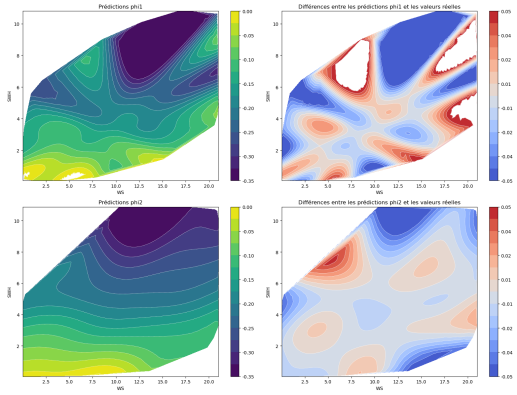
Figure 33: Results after 150 iterations for model φ_1 and φ_2 trained separately, $\alpha = 0.001$

Based on these initial results, we have decided to proceed with a single hidden layer consisting of 16 neurons for further experimentation and optimization.

The second test involves comparing the chosen architecture, where each model (φ_1 and φ_2) is trained separately for 150 iterations (Fig.34a), with an approach where the first model is trained for 50 iterations, followed by 50 iterations on the second model. After this initial training, we implement an alternating strategy where both models are optimized together with the loss function linked between them, continuing this process for an additional 50 iterations (Fig.34b.



(a) Models trained separately (150-150 iterations)

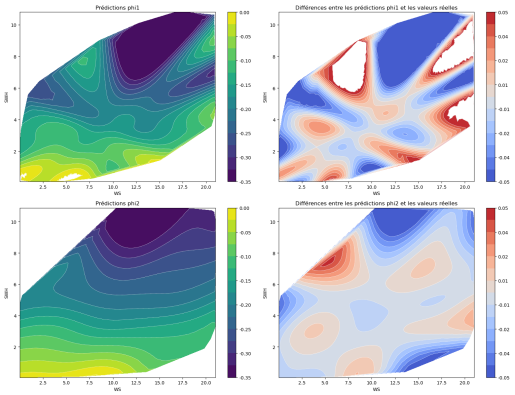


(b) Models trained (50-50 iterations) and alternating process (50 iterations)

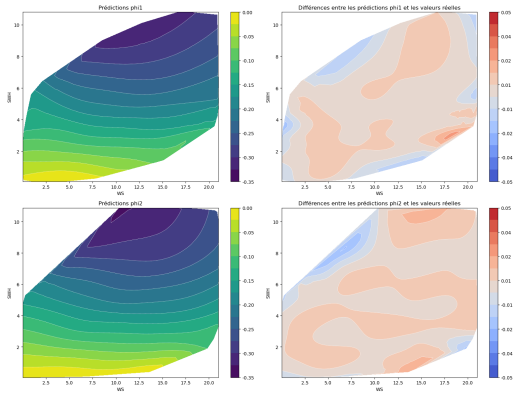
Figure 34: Results for model φ_1 and φ_2 trained only separately or also alternating, $\alpha = 0.001$

This approach significantly enhances the learning process, as evidenced by the reduction in differences between the predicted and theoretical models. This improvement is particularly noticeable in the right column, where the colors are lighter, indicating smaller errors and a closer alignment with the expected SSB values. Based on these results, we validate this architecture for further experiments.

The next test involves comparing the performance of the model with the original learning rate ($\alpha = 0.001$) to a higher learning rate ($\alpha = 0.05$). This will help us evaluate the impact of the learning rate on the convergence speed and accuracy of the model.



(a) Training with $\alpha = 0.001$

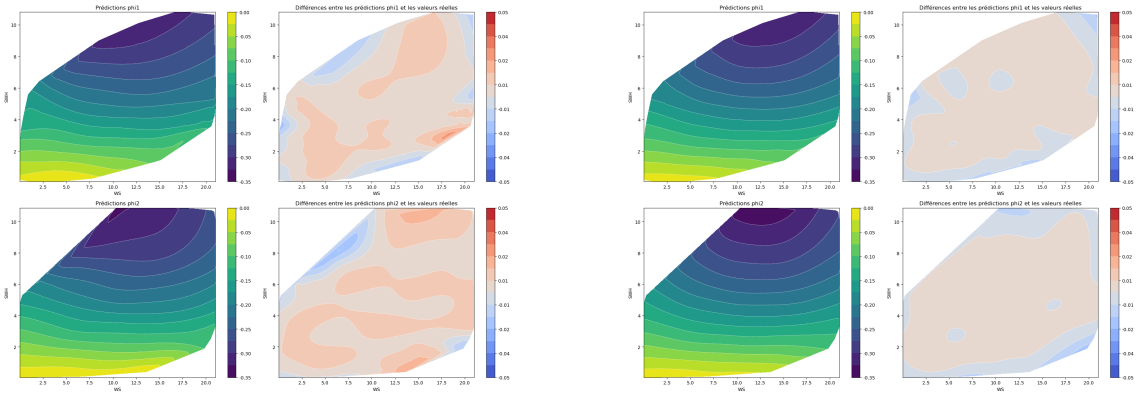


(b) Training with $\alpha = 0.05$

Figure 35: Results for model φ_1 and φ_2 trained separately+alternating, with different values of α

This approach leads to a significant improvement in the model's performance in terms of difference. Moreover, the predicted SSB structure closely aligns with the theoretical function, displaying the expected horizontal gradients. Based on these encouraging results, we have chosen this architecture for further development and analysis.

In this subsequent test, we refine our approach by allowing each phase of the training process to converge based on a predefined loss criterion (Fig.36b) rather than a fixed number of iterations (Fig.35a). Specifically, we set the threshold for the loss function at 10^{-6}, which corresponds to the expected precision of 1 cm for the SSB. The model progresses to the next phase only when this criterion is met, ensuring that each model is sufficiently trained before moving forward. This method aims to further enhance the accuracy of the final SSB predictions by allowing more dynamic and precise adjustments during the training process.



(a) Convergence after 50 iterations of each phase

(b) Convergence with $loss < 10^{-6}$ criteria

Figure 36: Results for model φ_1 and φ_2 trained separately+alternating, with different stop criteria, $\alpha = 0.05$

The new architecture took 17835 seconds to run for the training, with the training process divided into three distinct phases: 4 epochs for the convergence of φ_1 model, 27 epochs for the convergence of φ_2 model, and 3986 epochs for the final alternating phase.

This approach significantly improves the model's performance, resulting in differences between 0.01 and -0.01. Moreover, it keeps the horizontal gradients structure of the predicted SSB structure. Given these promising results, we select this architecture for further development and analysis.

We are very satisfied with this solution and have selected this model for the analysis using synthetic data: a hidden layer with 16 neurons, an alpha value of 0.05, nested models with alternating phases for the final training stage, and a convergence criterion on the loss set to 10^{-6}.

We also present the training loss curves for this model in Fig. 37, which show that the convergence stabilizes well, with the loss values observed on a scale of 10^{-3}.

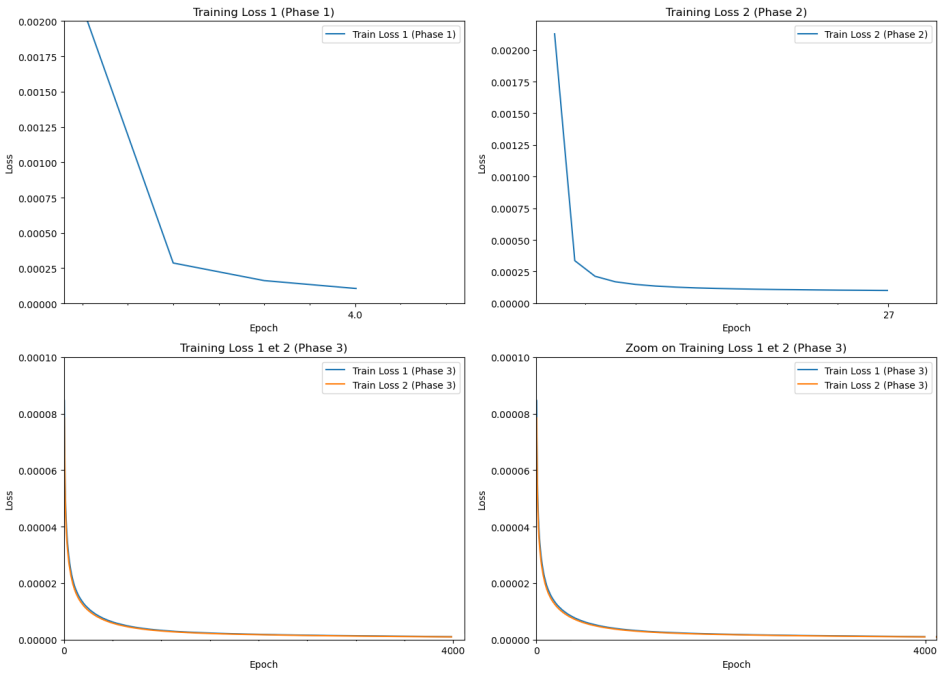


Figure 37: Loss curves for the 3 phases (model φ_1 loss in column 1, model φ_2 loss in column 2)

The results showed that the neural network model was able to replicate the BM4 parametric function to a high degree of accuracy. The 2D-function shape produced by the neural network closely matched the expected output from the BM4 model, indicating that the model was successfully learning the relationships between the input variables and the SSB estimations.

2. Transition to Real-World Data

The approach defined with synthetic data is then applied to real-world data, where the true complexity of the problem became apparent. The real-world data presented challenges due to the significant noise in the observed ΔSSH, and the task to extract the true SSB signal is difficult. Even with the tested parameters and adjustments, it proved challenging to achieve significant improvements, indicating the inherent complexity of the data and the limitations of the current approach.

4.3.5 Conclusion and Future Work

This study explored the application of neural networks to estimate SSB and the results demonstrate varying degrees of success. While the performance on synthetic data is satisfying, the transition to real-world data still presents challenges. However, we have already established a solid pipeline that will allow to refine and improve the model further with real data.

Key Findings

- Synthetic Data Success: The neural network model effectively learned the BM4 parametric function, demonstrating its ability to capture the key relationships between input variables under controlled conditions. The use of a simpler architecture with one hidden layer of 16 neurons, an alpha value of 0.05, and a loss convergence criterion of 10^{-6} was validated as the most effective configuration for this task.
- Hyperparameter Tuning: The results highlighted the importance of careful hyperparameter tuning. Adjustments of the learning rate, number of layers, and number of neurons had a significant impact on the model's performance, particularly when transitioning from synthetic to real-world data.
- Real-World Data Challenges: Transitioning to real-world data introduced significant noise, which complicated the model's ability to accurately predict SSB. While the model did not outperform traditional non-parametric methods on real data, it showed potential for handling higher-dimensional inputs and complex interactions.
- Comparison to Traditional Methods: The neural network approach did not yet achieve the same level of accuracy as the traditional non-parametric methods. However, its ability to model complex interactions suggests that with further refinement, it could become a powerful tool for SSB estimation.

Future Work

- Noise Reduction Techniques: Future efforts should focus on improving the model's robustness to noise in real-world data, potentially through advanced regularization techniques or data augmentation.
- Hybrid Approaches: Combining the strengths of neural networks with traditional methods, such as using neural networks to preprocess data before applying non-parametric smoothing, could enhance overall model performance.
- Extended Validation: Further validation of the neural network model with larger and more diverse datasets, as well as cross-validation to assess overfitting, will be critical in refining the approach and ensuring its generalizability.

In conclusion, while the neural network model demonstrated potential on synthetic data, further development is needed to fully realize its capabilities on real-world data. The established pipeline provides a solid foundation for ongoing research and optimization.

5 Feedback

This apprenticeship at CLS provided me with the opportunity to immerse myself in the field of altimetry and oceanography. It was quite challenging and I managed to leverage my data science and artificial intelligence skills and effectively integrate them with newfound expertise in applied altimetry. The work was difficult but very rewarding, as it allowed me to learn a wide range of new tools and techniques, including the use of PyTorch for neural network models, as well as the integration of Fortran and Python for the non-parametric model, and the insights into kernel smoothing methods. Interpreting physical and oceanographic data within this rigorous scientific framework has been an enriching experience, enhancing both my technical skills and my understanding of the complexities of oceanographic modeling.
In moving forward, I plan to refine my scheduling and task prioritization methods to streamline workflows. This will involve setting clearer milestones and deadlines, allocating adequate time for team interactions and feedback incorporation into the project cycle.
In the end, this apprenticeship has reinforced my desire for working on environmental issues. It has also opened doors to new applications for data science and further solidified my interest in cultivating concrete application to my work.

6 Conclusion

This study aimed to enhance the accuracy and effectiveness of Sea State Bias (SSB) modeling in satellite altimetry by exploring two innovative approaches: dimensionality reduction through Principal Component Analysis (PCA) combined with non-parametric kernel smoothing, and the direct use of neural networks to model SSB with five variables.

The PCA-based approach provided a simplified model by reducing the dimensionality of the dataset, allowing us to apply well-established 2D non-parametric methods. While the results showed that this method could approximate the SSB, it did not outperform the traditional models based on SWH and WS. This suggests that further optimization of kernel smoothing parameters and possibly the inclusion of additional principal components or alternative techniques could be beneficial.

On the other hand, the neural network approach showed significant potential, especially when applied to synthetic data, where it successfully learned the underlying relationships of the BM4 parametric model. However, when transitioning to real-world data, the neural network model faced challenges, primarily due to the noise present in the data. Despite this, the flexibility of neural networks to handle complex, high-dimensional data points to their future potential, especially with further refinements in handling noise and integrating hybrid approaches.

In summary, while the PCA and neural network methods have shown promise, more work is required to optimize these approaches fully. The study has established a strong foundation and a clear pathway for future research, with the goal of further improving SSB models and, ultimately, the precision of sea level measurements in satellite altimetry.

References

[1] N.Tran; L.Amarouche; P.Dubois; S.Labroue (CLS) D. Vandemark (UNH). *Altimétrie spatiale - Sea State Bias*. PowerPoint Presentation. Ecole d'Eté 2014, Saint-Pierre d'Oléron. Sept. 2014.

[2] Claire Dufau (CLS). *Introduction à l'océanographie*. PowerPoint Presentation, Oceanography Module. ISAE. Mar. 2011.

[3] Philippe Gaspar and Jean-Pierre Florens. "Estimation of the Sea State Bias in Radar Altimeter Measurements of Sea Level: Results from a New Nonparametric Method". In: *Journal of Geophysical Research: Oceans* 103.C8 (1998), pp. 15803–15814. DOI: 10.1029/98JC01265. URL: https://www.researchgate.net/publication/255700594_Estimation_of_the_sea_state_bias_in_radar_altimeter_measurements_of_sea_level_Results_from_a_new_nonparametric_method.

[4] Philippe Gaspar et al. "Improved Nonparametric Estimates of the Sea State Bias in Radar Altimeter Measurements of Sea Level". In: *Journal of Atmospheric and Oceanic Technology* 19.10 (2002), pp. 1690–1707. DOI: 10.1175/1520-0426(2002)019<1690:INEOTS>2.0.CO;2. URL: https://journals.ametsoc.org/view/journals/atot/19/10/1520-0426_2002_019_1690_ineots_2_0_co_2.xml.

[5] Rosemary Morrow (CTOH;LEGOS). *Satellite Altimetry*. PDF Document. Satellite Oceanography, CICESE, Lecture Notes. Sept. 2008. URL: https://www.aviso.altimetry.fr/fileadmin/documents/kiosque/education/Rose_cours1_2008.pdf.

[6] Ruoqing Zhu PhD. *Chapter 11 Kernel Smoothing — Statistical Learning and Machine Learning with R*. URL: https://teazrq.github.io/SMLR/kernel-smoothing.html (visited on 08/26/2024).

[7] *Sea Surface Heights*. GGOS. URL: https://ggos.org/item/sea-surface-heights/.

[8] *Significant Wave Height and H2*. Apr. 2016. URL: https://www.researchgate.net/post/Significant_Wave_height_and_H2.

[9] *Stokes Drift*. Wikipedia, the free encyclopedia. Oct. 2023. URL: https://en.wikipedia.org/wiki/Stokes_drift.

[10] Pierre Thibaut (CLS). *Satellite Altimetry Training*. PowerPoint Presentation, Altimetry Training Lessons. ENSAE. Mar. 2011.

[11] Ngan Tran et al. "Recent Advances in Satellite Altimetry for Earth Observation". In: *Advances in Space Research* 68.2 (2021), pp. 537–571. DOI: 10.1016/j.asr.2019.09.007. URL: https://www.sciencedirect.com/science/article/abs/pii/S0273117719308427.

[12] Ngan Tran et al. *The Sea State Bias in the Jason-2 Altimetry*. Tech. rep. Centre National d'Études Spatiales (CNES), 2010. URL: https://archimer.ifremer.fr/doc/00002/11276/7851.pdf.

	Global		**North Pacific**		**Equator**		**South Pacific**	
Variable	PC1	PC2	PC1	PC2	PC1	PC2	PC1	PC2
Orbital velocity	**0.954**	-0.205	**0.810**	-0.171	**0.451**	0.198	**0.927**	-0.204
Stokes Drifts	0.567	-0.015	0.125	0.164	-0.171	-0.026	0.706	0.129
Mean Wave Period	0.550	**0.822**	0.414	**0.775**	-0.325	**0.588**	0.388	**0.788**
Wind Speed	**0.836**	-0.500	**0.745**	-0.446	**0.577**	0.045	**0.837**	-0.462
SWH	**0.956**	0.178	**0.776**	0.167	0.230	0.310	**0.928**	0.193
	NW Pacific		**SE Pacific**		**Indian Ocean**		**Wide Tropics**	
Variable	PC1	PC2	PC1	PC2	PC1	PC2	PC1	PC2
Orbital velocity	**0.994**	0.144	**0.929**	-0.202	**0.720**	-0.149	0.368	0.368
Stokes Drifts	0.084	**-0.735**	0.781	0.216	-0.071	0.009	-0.173	-0.073
Mean Wave Period	0.642	-0.490	0.237	**0.804**	0.358	**0.824**	**-0.607**	**0.519**
Wind Speed	**0.902**	0.366	**0.873**	-0.441	**0.651**	-0.393	**0.545**	0.240
SWH	**0.987**	-0.098	**0.919**	0.232	**0.626**	0.111	0.085	0.427

Table 4: Loading coefficients for the first two principal components for Dataset 1 (SAR) across different geographical areas

	Global		**North Pacific**		**Equator**		**South Pacific**	
Variable	PC1	PC2	PC1	PC2	PC1	PC2	PC1	PC2
Orbital velocity	**0.961**	-0.196	**0.817**	-0.173	**0.466**	0.187	**0.933**	-0.190
Stokes Drifts	0.573	-0.019	0.131	0.150	-0.179	-0.021	0.708	0.144
Mean Wave Period	0.542	**0.832**	0.412	**0.780**	-0.307	**0.599**	0.373	**0.797**
Wind Speed	**0.845**	-0.495	**0.736**	-0.446	**0.586**	0.044	**0.864**	-0.458
SWH	**0.968**	0.173	**0.769**	0.167	0.216	0.312	**0.934**	0.186
	NW Pacific		**SE Pacific**		**Indian Ocean**		**Wide Tropics**	
Variable	PC1	PC2	PC1	PC2	PC1	PC2	PC1	PC2
Orbital velocity	**1.001**	0.165	**0.935**	-0.183	**0.730**	-0.144	0.383	0.363
Stokes Drifts	0.091	**-0.675**	0.782	0.240	-0.076	0.009	-0.180	-0.071
Mean Wave Period	0.641	**-0.542**	0.218	**0.809**	0.345	**0.831**	**-0.593**	**0.538**
Wind Speed	**0.858**	0.396	**0.898**	-0.430	**0.681**	-0.387	**0.552**	0.244
SWH	**0.987**	-0.097	**0.910**	0.213	**0.599**	0.137	0.076	0.423

Table 5: Loading coefficients for the first two principal components for Dataset 2 (ERA-5) across different geographical areas

	Global		**North Pacific**		**Equator**		**South Pacific**	
Variable	PC1	PC2	PC1	PC2	PC1	PC2	PC1	PC2
Orbital velocity	**0.950**	-0.214	**0.806**	-0.181	**0.446**	0.203	**0.923**	-0.212
Stokes Drifts	0.568	-0.030	0.127	0.163	-0.169	-0.028	0.707	0.113
Mean Wave Period	0.560	**0.814**	0.425	**0.766**	-0.329	**0.584**	0.397	**0.783**
Wind Speed	**0.831**	-0.499	**0.744**	-0.451	**0.585**	0.058	**0.830**	-0.466
SWH	**0.951**	0.188	**0.766**	0.176	0.202	0.313	**0.928**	0.207
	NW Pacific		**SE Pacific**		**Indian Ocean**		**Wide Tropics**	
Variable	PC1	PC2	PC1	PC2	PC1	PC2	PC1	PC2
Orbital velocity	**0.991**	0.127	**0.926**	-0.211	**0.717**	-0.153	0.354	0.379
Stokes Drifts	0.084	**-0.820**	0.783	0.193	-0.070	0.007	-0.170	-0.079
Mean Wave Period	0.650	-0.392	0.248	**0.803**	0.366	**0.819**	**-0.622**	**0.500**
Wind Speed	**0.903**	0.305	**0.865**	-0.444	**0.660**	-0.398	**0.538**	0.266
SWH	**0.979**	-0.080	**0.920**	0.249	**0.603**	0.123	0.048	**0.423**

Table 6: Loading coefficients for the first two principal components for Dataset 3 (LRM) across different geographical areas

Proof with the geometric formulation of the first order condition for minimization

- **Reminder:**

For multiple linear regression of the type $y = \beta X + \epsilon$, where y is the vector of observations, X is the matrix of explanatory variables, β is the vector of parameters to estimate, and ϵ is the vector of errors, we seek to estimate $\hat{\beta}$, the least squares estimator of β.

The objective is to minimize the sum of the squares of the residuals, that is, to find $\hat{\beta}$ that minimizes the following cost function:

$$\min_{\beta} \|y - \beta X\|^2$$

This minimization is equivalent to searching for the orthogonal projection of the vector y onto the space spanned by X, denoted $[X]$. The orthogonal projection ensures that the distance between y and the prediction βX is minimal.

Mathematically, for $y - \hat{\beta}X$ to be orthogonal to the space $[X]$, the dot product between $y - \hat{\beta}X$ and any linear combination of the columns of X (denoted $\tilde{X}$) must be zero:

$$\forall \tilde{X}, \langle y - \hat{\beta}X, \tilde{X} \rangle = 0$$

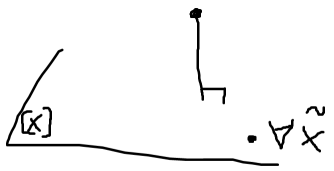


Figure 38: Vector space spanned by X and orthogonal projection

Since $\tilde{X}$ can be represented by any linear combination of X, let's replace it with $\tilde{\beta}X$. We have:

$$\forall \tilde{\beta}, \langle y - \hat{\beta}X, \tilde{\beta}X \rangle = 0$$

Hence:

$$\forall \tilde{\beta}, \tilde{\beta}X^T(y - \hat{\beta}X) = 0$$

Thus:

$$\forall\tilde{\beta}, \tilde{\beta}(X^T y - \hat{\beta}X^T X) = 0$$

However, the orthogonal of a linear form being 0, we have:

$$X^T y - \hat{\beta}X^T X = 0$$

By isolating $\hat{\beta}$, we find the least squares solution for $\hat{\beta}$:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

This shows that $\hat{\beta}$ is obtained through the orthogonal projection of y onto the space spanned by X, and this solution minimizes the sum of the squares of the residuals between the observations y and the predictions $\hat{\beta}X$.

- **In our case**, we aim to solve

$$\min_{\varphi_1, \varphi_2} \|z - (\varphi_2(X_2) - \varphi_1(X_1))\|^2$$

We orthogonally project z onto the space spanned by $(\varphi_2 - \varphi_1)$, so $z - (\varphi_2 - \varphi_1)$ will be orthogonal to $(\tilde{\varphi_2} - \tilde{\varphi_1})$.
Thus, for all functions $\tilde{\varphi_2}, \tilde{\varphi_1}$, we have:

$$E[(z - (\varphi_2 - \varphi_1))(\tilde{\varphi_2} - \tilde{\varphi_1})] = 0$$

This is equivalent to:

$$E[\tilde{\varphi_2}(z - (\varphi_2 - \varphi_1))] = E[\tilde{\varphi_1}(z - (\varphi_2 - \varphi_1))]$$

Under the assumption of exchangeability, X_1 and X_2 have the same marginal distribution (i.e., $\varphi_2 = \varphi_1 = \varphi$, $\tilde{\varphi_2} = \tilde{\varphi_1} = \tilde{\varphi}$), we have:

$$E[\tilde{\varphi}(X_2)(z - (\varphi(X_2) - \varphi(X_1)))] = E[\tilde{\varphi}(X_1)(z - (\varphi(X_2) - \varphi(X_1)))], \forall\tilde{\varphi}$$

Since this holds for all $\tilde{\varphi}$, if we take a function $\tilde{\varphi}$ that is zero for X_1 and non-zero for X_2, we then have the following expression:

$$E[\tilde{\varphi}(X_2)(z - (\varphi(X_2) - \varphi(X_1)))] = 0$$
$$E[\tilde{\varphi}(X_1)(z - (\varphi(X_2) - \varphi(X_1)))] = 0$$

Therefore, if this is true for all $\tilde{\varphi}$, then with the first equation:

$$E[\tilde{\varphi}(X_2)(E[z|X_2] - \varphi(X_2) + E[\varphi(X_1)|X_2])] = 0$$

Hence, if a certain quantity is orthogonal to all functions $\tilde{\varphi}$, then this quantity must be zero. This is an application of the idea that if the expectation of the product of a quantity with any function of the space is always zero, then this quantity is zero:

$$E[z|X_2] - \varphi(X_2) + E[\varphi(X_1)|X_2] = 0$$

Thus, we obtain:

$$E[z|X_2 = x] = \varphi(x) - E[\varphi(X_1)|X_2 = x]$$