Ocean Modelling 139 (2019) 101404

Contents lists available at ScienceDirect

Ocean Modelling

journal homepage: www.elsevier.com/locate/ocemod

A novel approach to computing super observations for probabilistic wave model validation

Patrik Bohlinger[a,*], Øyvind Breivik[a,c], Theodoros Economou[b], Malte Müller[a]

[a] *Norwegian Meteorological Institute, Allegaten 70, Postboks 7800, 5020 Bergen, Norway*
[b] *University of Exeter, UK*
[c] *University of Bergen, Norway*

ARTICLE INFO

Keywords:
Wave model
Validation
Super observation
Gaussian process
Machine learning

ABSTRACT

In the field of wave model validation, the use of super observations is a common strategy to smooth satellite observations and match the simulated spatiotemporal scales. An approach based on averaging along track is widely applied because it is straightforward to implement and adjustable. However, the choice of an appropriate length scale for obtaining the averages can be ambiguous, affecting subsequent analyses. Despite this dilemma, no uncertainty for the validation metric is provided when proceeding with wave model validation. We show that super observations computed from averaging data points applying an inappropriate length scale can lead to a misrepresentation of the wave field which can introduce errors into the wave model validation. Modelling the mean of observations as a Gaussian Process mitigates those errors and reliably identifies outliers by exploiting information hidden in the observational time series. Moreover, the uncertainty accompanying the validation statistic is readily accessible in the Gaussian Process framework. The flexibility of a Gaussian process makes it an attractive candidate for the probabilistic validation of wave models with steadily increasing horizontal resolution. Moreover, this approach can be applied to measurements from other platforms (e.g. buoys) and other variables (e.g. wind).

1. Introduction

Satellite altimetry allows one to evaluate the performance of wave models in terms of significant wave height (SWH). Multiple satellite missions in the past have created an almost continuous record since 1985, allowing for wave model validation and calibration by measurement platforms (Zieger et al., 2009; Abdalla et al., 2011). However, prior to validation, the satellite derived SWH undergoes quality control, i.e. outlier removal, and is adjusted to match the effective horizontal resolution of the wave model (Abdalla et al., 2013). The resulting values are often called super observations. We propose a new and highly flexible approach to compute super observations, detect outliers, and attain a probabilistic uncertainty measure for subsequently computed validation statistics.

Commonly, a super observation is computed by averaging along the satellite track of SWH (an example of which is shown in Fig. 2 a), as is repeatedly applied and described in recent scientific literature (Abdalla et al., 2011; Liu et al., 2016; Stopa et al., 2016). The size of the averaging window (length scale) is chosen to match the effective model resolution, such that values representing similar process scales are compared (Abdalla et al., 2013). Another reason for the smoothing is the chaotic nature of the sea state consisting of a superposition of multiple wave fields with different origin and characteristics (random phase/amplitude model, e.g. Holthuijsen (2010)). From an Eulerian perspective, this superposition creates a varying wave field with random realizations scattered around a mean value. This mean value can be thought of as the sea state one strives to measure but can rarely observe. As one satellite footprint is just a snapshot of the wave field at one point in space and time, it is unlikely that it recorded exactly the underlying sea state but rather one random realization. Smoothing the time series of along track measurements using an appropriate length scale averages out these variations and hopefully arrives at the correct mean. Averaging will further reduce the effect of random measurement errors inherent to the measuring system.

A challenge with the described approach is the choice of an appropriate length scale. Even though guidelines are provided, e.g. by Abdalla et al. (2011), one has to decide on one specific value for the length scale. This is commonly the mean of what might be the range of possible length scales. However, the most appropriate length scale might vary in space, or from domain to domain, because it depends on

[*] Corresponding author.
E-mail address: patrikb@met.no (P. Bohlinger).

https://doi.org/10.1016/j.ocemod.2019.101404
Received 4 February 2019; Received in revised form 16 May 2019; Accepted 18 June 2019
Available online 22 June 2019
1463-5003/

the scale of features the wave model is able to resolve. Unfortunately, in the prevalent approach, one can only choose one discrete number of footprints or model grid cells that should be averaged.

If the correct length scale could be chosen and there is no systematic bias in the data, the super-observed average SWH value is a reliable measure, assuming the sample size (averaging window size) is large enough to be representative of the underlying stochastic process. This assumption is challenged by model simulations with increasingly higher horizontal resolution. For instance, when comparing satellite observations against high resolution model simulations, the appropriate number of footprints that should be averaged might reduce to only a few points. Abdalla et al. (2011) chose 11 consecutive valid footprints to form one super observation and Saleh Abdalla (2018) mentioned a necessary minimum of 7 footprints. What if the resolved model scale drops below that value and what if the 11 or 7 values are not representative? This becomes increasingly problematic considering an ever-greater mismatch between the model resolution and the size and distance of the satellite footprints. Smaller numbers are already applied, e.g. Stopa et al. (2016) used a moving average window of 5 values to smooth the satellite track and to match scales.

A similar problem emerges for the detection of outliers. As described e.g. by Young (1999) or Zieger et al. (2009), outliers are detected based on a block of data points (25 values in Zieger et al. (2009)) which undergoes 2 or 3 passes of quality checks. Again, the block size has to be decided with regard to physical arguments like the spatial scale of geophysical processes which, however, can vary in space and time. The assumption is that the block size should be representative for computing a mean and a standard deviation but small enough to avoid considerable variability as a result of e.g. storm systems. This is not the case if there are strong gradients or regime shifts along the satellite track within the chosen block.

After a sequence of subjective decisions on the length scale and block size, a validation statistic can be computed. Commonly single numbers are presented, e.g. root mean square error, correlation, or scatter index. However, the uncertainty associated with these decisions is not quantified and as such, the significance of the validation statistics cannot be realistically assessed. Propagating and quantifying this uncertainty is one of the primary concerns in the exposition of the approach in this paper.

In the following, we present a flexible, alternative approach which allows us to estimate the most likely sea state along the satellite track and thus to create reliable and smooth super observations detecting outliers at the same time. Moreover, our suggested approach can naturally quantify the uncertainty to the desired validation statistic.

2. Methodology

2.1. Data and collocation

For producing the results in this paper, we used two data sources: First, output from simulations with the operational wave model Arctic WAM, a version of the WAM wave model (Komen et al., 1994), setup with a horizontal resolution of 8 km regridded to 6.25 km. The simulations were conducted at the Norwegian Meteorological Institute and are available on the Copernicus web server under the product name ARCTIC_ANALYSIS_FORECAST_WAV_002_010.[1] Second, Sentinel-3a (S3a) data obtained from the Copernicus web server under the product name WAVE_GLO_WAV_L3_SWH_NRT_OBSERVATIONS_014_001.[2] S3a is a level three satellite altimeter product featuring a 1 Hz sampling frequency. The wave model is forced with winds from the operational ECMWF IFS at 9 km horizontal resolution and at the boundaries with the operational ECMWF wave model with a horizontal resolution of 14 km. A more thorough description is provided on the web site for these Copernicus products.

To obtain a time series from the wave model matching the satellite observations, we collocate the S3a footprints to model grid cells (Fig. 1). Constraints in space and time determine whether a model grid cell is attributed to a satellite footprint comparable to Stopa et al. (2016). The time constraint is centered around the model time step and allows S3a values to be chosen if they were recorded within ± 30 min. In space, we allow only the collocation of model grid cells which are directly associated with the satellite swath with a maximum distance of 6 km to the footprint. This approach results in two time series, one for the wave model and one for the satellite observations, consisting of the same number of values.

2.2. Averaging based super observations and outlier detection using data blocks

We compute an effective length scale from the horizontal resolution of the atmospheric model by multiplying 9 km by 3–6 grid cells as described in Abdalla et al. (2011). The wave model has a higher resolution and is thus not the limiting factor. Our length scale is consequently $l = 3 \cdot 9$ km $= 27$ km to $l = 6 \cdot 9$ km $= 54$ km consistent with 3–7 wave model grids and 4–8 consecutive satellite footprints. Since we need an odd number for a centered moving average window, we tested both 5 and 7 satellite footprints. For the illustrations we show only results with window size 7. We acknowledge that in practice, often independent blocks are used rather than a moving average (Saleh Abdalla, 2018, personal communication). The independent block estimates are points on the smoothed line resulting from the moving average, and as such are implicitly included in our comparison. For the sake of illustration and because this is sometimes applied, we chose to compare our method against moving average based super observations.

Outlier removal is performed prior to computing super observations. This was approached by dividing the time series into smaller blocks of retrieved altimeter SWH values and removing values greater than twice the standard deviation from the block mean. Tobe consistent with the computation of the super observations and the effective model scale, we compute outliers based on the same length scale of 7 consecutive observations. Outliers are additionally detected based on larger block sizes of 11 and 25 values for comparison with Abdalla et al. (2011) and Zieger et al. (2009).

2.3. Super observations and outlier detection with a Gaussian process model

The description and application of a Gaussian Process is based on Rasmussen and Williams (2006) who introduce Gaussian Processes for machine learning. We assume that an unknown stochastic process, non-stationary in time and space, produces a wave field. For the short amount of time that the satellite needs to pass over the model region, typically some minutes, we assume that the stochastic process is stationary. This assumption is realistic because the sea state does not noticeably change at the location of the footprints within this amount of time. This means that the SWH along any trajectory over the wave field is a function of space, or similarly travel time of the footprint. For any point in space in the wave field there is a true mean value of the sea state around which measurements would scatter due to the chaotic nature of the sea and random measurement errors. The true sea state can be thought of as a latent mean function of SWH we wish to estimate and subsequently use for validation of the wave model.

2.3.1. Description of our statistical model

From a statistical point of view, the SWH time series can be

[1] http://marine.copernicus.eu/services-portfolio/access-to-products/?option=com_csw&view=details&product_id=ARCTIC_ANALYSIS_FORECAST_WAV_002_010.

[2] http://marine.copernicus.eu/services-portfolio/access-to-products/?option=com_csw&view=details&product_id=WAVE_GLO_WAV_L3_SWH_NRT_OBSERVATIONS_014_001.

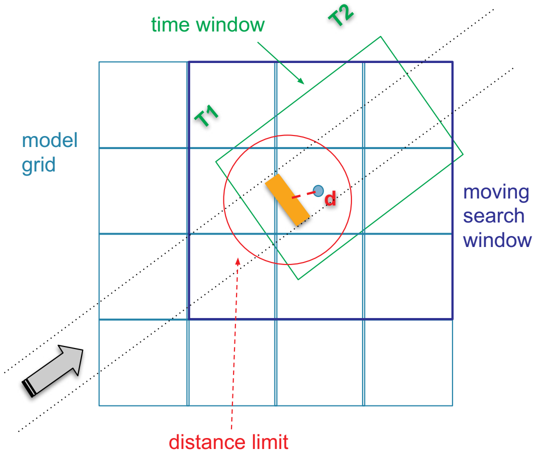

Fig. 1. Sketch of the used collocation method. The satellite footprint is illustrated with an orange rectangle. For speedup collocation only takes place within a search window (indicated in blue) where all model values that are not in the vicinity are masked out.

described as a mean (trend) plus error model, where the trend is what we are after, while the error captures the aggregated effect from measurement error and natural variability in the wave field. This model needs to have a flexible enough trend to capture characteristics of the true mean SWH, such as gradients, turning points and local maxima/minima. Here, we take an approach where the trend is characterized non-parametrically (as opposed to parametrically, e.g. linear, quadratic etc.) using a Gaussian Process (GP). This assumes that the trend is an unknown smoothly varying function of time, that can be estimated from the available data. Features leaving their imprint on the time series can be e.g. the spatial extent of weather patterns of different scales, wind shadow effects due to topography, shelter effects due to land, or attenuation effects due to sea ice. Due to the dependency of the data points, much of the necessary information lies in the measured values and their distance to each other.

Denote an observation of (mean-centered) SWH at time t by y_t. We consider the following model:

$$y_t|f_t = f_t + \epsilon_t \tag{1}$$

$$\epsilon_t \sim N(0, \sigma_n^2) \tag{2}$$

so that $y_t\,|\,f_t \sim N(f_t, {\sigma_n}^2)$. This model assumes that conditional upon the trend f_t, the error about this trend is independent Gaussian noise with variance ${\sigma_n}^2$ (where n denotes the total number of time points in the data). Furthermore, the trend itself f_t is also assumed to be Gaussian, but not one that is independent in time. Values f_t and $f_{t'}$ will be assumed positively correlated with the strength of correlation decreasing with the temporal separation $|t - t'|$. Specifically, f_t is assumed to have zero mean and variance ${\sigma_s}^2$:

$$f_t \sim N(0, \sigma_s^2) \tag{3}$$

and the dependence f_t and $f_{t'}$ defined by the squared exponential correlation function:

$$\mathrm{cor}(f_t, f_{t'}) = \exp\left(-\frac{(t-t')^2}{2l^2}\right) \tag{4}$$

This is a Gaussian Process (GP) with zero mean, variance parameter ${\sigma_s}^2$ and length scale parameter l. The correlation function is defined in such a way to ensure that the correlation decays with increasing temporal separation, and notice that this is squared in the exponential term ensuring a very smooth function f_t (infinitely differentiable), an expected behavior of a wave field. The length scale parameter (to be estimated from the data) controls the amount of smoothness, with small l resulting in wiggly looking functions whereas large l result in more slowly varying ones.

2.3.2. *Model estimation*

Notice that the probability distribution for y_t is defined conditionally on f_t. It turns out (Rasmussen and Williams, 2006) that the marginal probability distribution for the vector $y = (y_1, \ldots, y_n)$ is a multivariate Gaussian distribution $y \sim N(0, \Sigma)$, with mean vector zero and covariance matrix $\Sigma = K + {\sigma_n}^2 I$, where K is the covariance matrix that results from the GP and I is an $n \times n$ identity matrix that adds the Gaussiannoise. The diagonal entries of K are all equal to ${\sigma_n}^2 + {\sigma_s}^2$ whereas the off diagonals capture the covariances between all combinations of y_t and y_t' for $t \neq t'$, and are defined by

$$y_t, y_{t'} = \sigma_s^2 \exp\left(-\frac{(t-t')^2}{2l^2}\right) \tag{5}$$

We can now write down the likelihood of the data as the probability density function of the multivariate Gaussian distribution (Rasmussen and Williams, 2006) and proceed to maximize it with respect to the parameters ${\sigma_n}^2$, ${\sigma_s}^2$ and l. This will provide optimal point estimates for the three parameters, in accordance to likelihood theory. The algorithm for computing the latent mean function, its variance and the likelihood is described in Rasmussen and Williams (2006) (algorithm 2.1). For computational efficiency, we used the scikit-learn python package (Pedregosa et al., 2011) for optimization.

2.3.3. *Outlier detection*

In the GP framework, values can be flagged as outliers if they fall outside of the e.g. 2σ bounds computed in the previous section. This can be adjusted according to the desired threshold applying the algorithm described in Section 2.4.1. We apply the 2σ threshold to produce a result comparable with the prevalent approach.

2.4. *Uncertainty and probabilistic validation*

We estimate uncertainties of the validation statistics in two different ways, applying parametric bootstrap on the GP model and non-parametric bootstrapping on the unprocessed (raw) observational data. Both approaches lead to a large number of realizations of (super) observational time series which can be validated against the wave model time series. This results in a distribution of the computed validation statistic rather than a single value.

2.4.1. *Uncertainty estimation for the GP model*

To quantify and propagate the uncertainty involved in estimating the three parameters (σ_n,σ_s and l), we use parametric bootstrapping (Davison and Hinkley, 1997). This involves simulating data points ${y_t}^*$ that correspond toplausible realizations of the actual observations y_t, under the GP model.

First, it is instructive to consider the probability distribution of the trend f_t given the model and the data y_t. Denote this as $f_t\,|\,y_t$, something that can be interpreted as the best guess of the (smooth) trend given that we have observed some data. It turns out that this distribution is a multivariate Gaussian with a mean vector and covariance matrix given in Rasmussen and Williams (2006) (Chapter 2.2). The expressions for both the mean and covariance depend on the three model parameters (${\sigma_n}^2$, ${\sigma_s}^2$, l) as well as the data. Given the estimates of the three parameters obtained in Section 2.3.2, this distribution captures two sources of uncertainty: the stochastic variability assumed by modelling f_t as a GP and the sampling uncertainty in only having observed a finite sample of n data points. Notice however that this does not include the uncertainty involved in the estimation of the three model parameters. In what follows, we describe a simulation approach to quantify this appropriately.

The parametric bootstrapping algorithm is effectively a loop with index i that involves the following steps:

1. From the multivariate Gaussian distribution of $f_t|y_t$, simulate a realization of the trend, $f_1^{(i)}, \ldots, f_n^{(i)}$;
2. Then simulate n replicate data points from the conditional distribution of y_t, namely $y_t^{(i)} \sim N(f_t^{(i)}, \sigma_n^2)$, using the estimate of σ_n^2 obtained in Section 2.3.2;
3. Fit the GP model again to the replicate data set $y_1^{(i)}, \ldots, y_n^{(i)}$, to obtain new parameter estimates $\sigma_n^{2(i)}$, $\sigma_s^{2(i)}$ and $l^{(i)}$;
4. Using the new estimated parameters, compute the trend as the mean of the multivariate Gaussian distribution of $f_t|y_t$, to obtain a new realization of the trend, $f_1^{*(i)}, \ldots, f_n^{*(i)}$.

Repeating these steps for a large number of times, e.g. $i = 1\ldots, m$ with $m = 1000$, will provide m realizations of the trend. These realizations are then approximate samples from a distribution that also captures the uncertainty in the estimation of σ_n^2, σ_s^2 and l. The mean of these realizations is still a best estimate of the trend given the data, an example of which is shown in Fig. 4. The variability (through the m samples) can be expressed by taking e.g. the 2.5% and 97.5% empirical quantiles as an estimate of the 95% confidence interval. In addition, the uncertainty expressed by these samples, can be propagated as required. For example, the validation statistic can be computed for each sample yielding a distribution of the validation statistic. This distribution will express the uncertainty in having to estimate the trend and also the three model parameters from a sample of data.

2.4.2. Uncertainty estimation for the raw observational time series

The standard non-parametric bootstrap (Efron, 1979) samples from the collocated observational time series. The observations are chosen above the wave model because we assume the observations to follow a stochastic process whereas the numerical wave model produces deterministic results, namely the significant wave height computed from the zeroth-order moment of the variance density spectrum (e.g. Holthuijsen (2010)). Moreover, the variance is noticeably higher in the un-smoothed observations creating a larger spread and should therefore describe the uncertainty more adequately. The observational time series is re-sampled 1000 times, ultimately resulting in a distribution of the validation statistic. This provides a sense of uncertainty for the respective measure based on the degree of scattering in the measurement.

3. Results

Our study is based on a S3a track around the model time step of 2018-05-02 00:00 UTC. This is an arbitrary choice and our conclusion is also valid for other time steps. Nonetheless, this track depicts crucial features which will be discussed in the following.

The chosen S3a track enters our domain in the North Atlantic heading toward Greenland, crosses Greenland, the Arctic and Russia, and continues from the Bering Sea to the North Pacific (Fig. 2 a). The satellite track is obviously divided into two parts where we will call the footprints in the North Atlantic *Track 1* and the footprints in the North Pacific *Track 2*.

In Track 1, the SWH increases toward Greenland and drops abruptly from ca. 5 m to ca. 1 m SWH within a distance of 380 km. This results in a spiky local maximum and a steep gradient in the time series (Fig. 3 a and c). The reason for the abrupt changes in SWH is the sheltering effect of Greenland. The winds are mostly from west due to a low pressure system that travels from west to east over Greenland (Fig. 2 b). Waves generated by the low pressure system are stopped by the southern tip of Greenland. This leads to a short effective fetch and consequently a young, wind generated wave field in the Irminger Sea close to the south-eastern Greenland coast. Strong gradients in winds along the southern Greenland tip are not rare and are known to the scientific community (Doyle and Shapiro, 1999) with significant effect on the hydrography (Pickart et al., 2003) and, as evident here, the wave field (Fig. 3 a). This is therefore a good example for a steep gradient and local maximum in the wave field as these phenomena are not rare and the emerging features need to be accounted for.

Track 2 is characterized by a gradual increase in SWH (Fig. 3 b). The increase of wave height toward the south is related to the increasing distance between the observation location and the coastline of the Kamchatka peninsula, under the prevailing westerly winds that occurred around the chosen time step (not shown). The time series depicts multiple suspicious values that attract attention by their isolated and pronounced deviation from their neighbors (Fig.3 c). The detection of outliers will be elucidated in the following.

3.1. Outlier detection

We focus on Track 2 because there are obvious outliers which should be handled correctly by an automated outlier detection algorithm. We apply the algorithm described in Section 2.2 to detect outliers based on block sizes of 7, 11, and 25 consecutive valid values. Fig. 4 a) illustrates the difference among super observations derived from different window sizes and the according difference between the catchment area for accepted values (shaded area). The choice of the block size affects considerably the detection of outliers. None of the chosen window sizes detects all of the values that are seemingly erroneous.

In contrast, the GP based 2σ confidence intervals is much smoother (Fig. 4 b) and the block size did not have to be chosen manually. Learning from the entire available time series of satellite observations, the excursion of the super observation time series due to the outliers

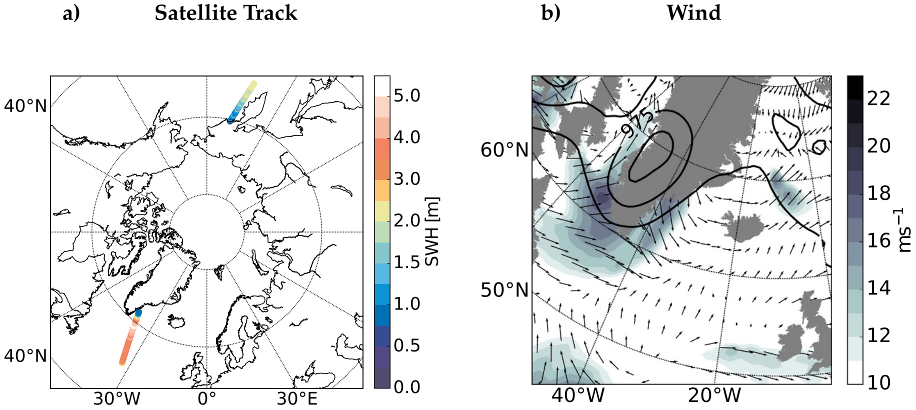

Fig. 2. a) S3a track according to the model time step at 2018-05-02 00:00 UTC. b) Wind speed and direction from 2018-05-01 12:00 UTC depicted in filled contours and arrows, respectively. Mean sea level pressure is shown in black contour lines with steps of 5 hPa.

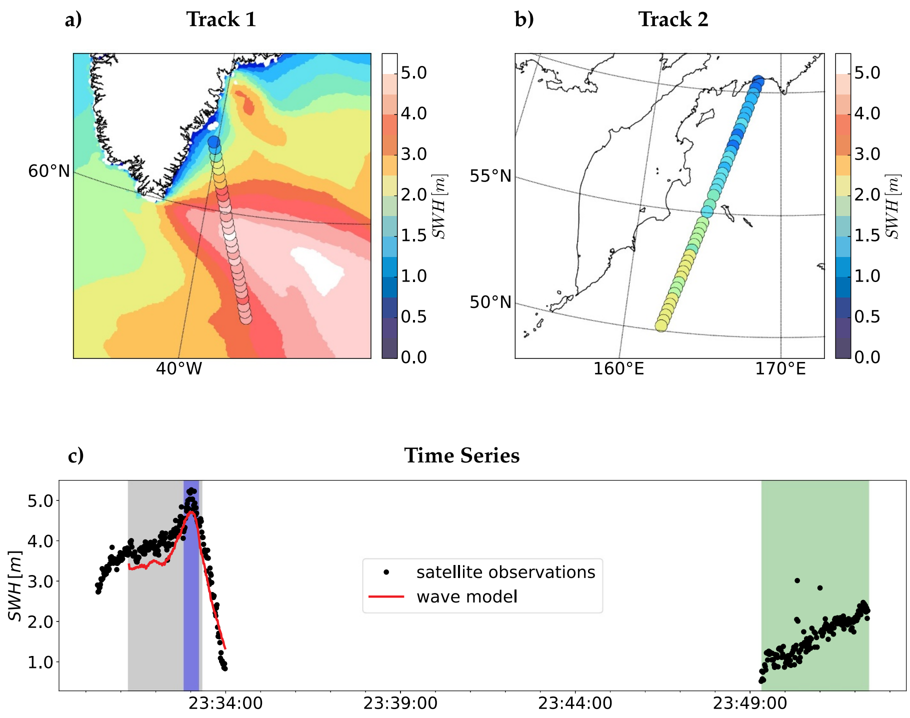

Fig. 3. a) S3a SWH depicted as round markers. Only every fourth value is displayed for readability. The filled contours are wave model derived SWH for the model output time 2018-05-02 00:00 UTC. b) same as a) but outside of the wave model domain. c) Time series of SHW from S3a (black dots) and the wave model red line. The red line represents only the collocated values of the wave model. The shaded areas depict sections of the time series that will be discussed in the results.

and the according 2σ region are more moderate. All values that seem erroneous lie outside the catchment region for acceptable values and can thus be flagged as outliers.

The aforementioned suspicious values were recorded when the satellite swath passed Bering Island. Erroneous recordings are common close to land-sea boundaries and near islands due to corrupted wave forms. This is why e.g. Young (1999) and Zieger et al. (2009) use land-sea masks for the detection of erroneous recordings. However, the suspicious footprints in our case were located over sea and several tens of km away from the coastline. This means that a land-sea mask would not help to detect these error seven if flagging values additionally in the vicinity of land.

A possible problem for an outlier detection algorithm is non-stationarity and non-linear behavior in the process that produces the stochastic data. Therefore, Zieger et al. (2009) assumed that within the block size of 25 values, representing a distance of ca. 180 km, the recordings of the sea state can be considered representative without significant variability due to geophysical processes. As illustrated in Fig. 3, this assumption can frequently be violated depending on the region of interest.

An example of such a violation is the local maximum close to the southern tip of Greenland (blue shading in Fig. 3 c). Fig. 5 zooms in on the local maximum showing a sample of recordings with a block size of 23 values. If this block had been chosen to detect outliers the leftmost value would have fallen outside of the catchment area for acceptable values and would have consequently been misinterpreted as an outlier. In a time series this behavior is difficult to anticipate and thus not straightforward to employ in an automatic outlier detection algorithm. However, when using a GP based approach the arc-like structure is automatically recognized and the sixth value from the right falls outside the blue shading. This seems more intuitive to the eye than the first value from the left.

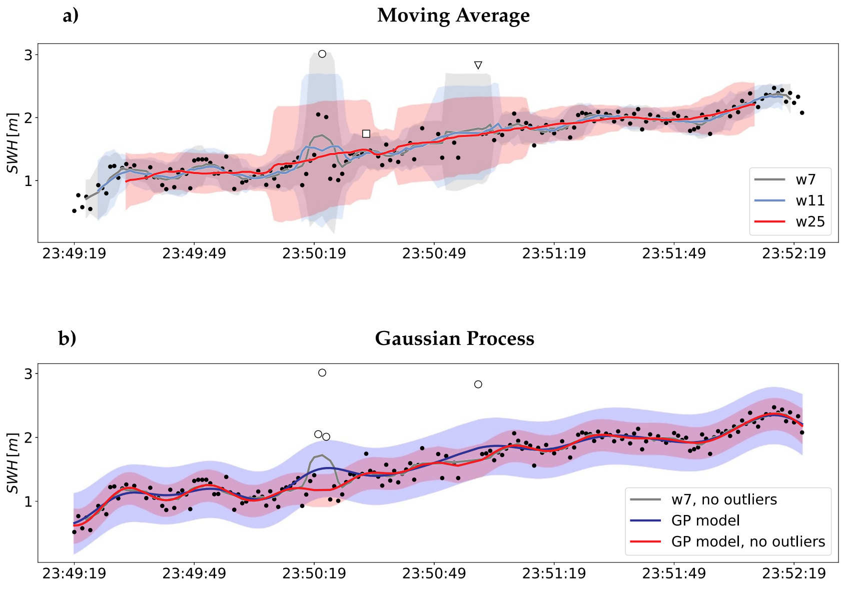

Fig. 4. Satellite footprints of Track 2 as black points together with outliers and smoothed time series. a) centered moving average with window sizes 7, 11, 25 as lines and 2σ interval in same color but shaded to mark the catchment area for acceptable values. Outliers are marked with a circle (detected with window size 25 and 11), a square (detected with window size 11 and 7) and a triangle (detected with all window sizes). b) Same as a) but with a moving average of window size 7 applied on the time series after removing outliers. The outlier detection is based on the same window size. Additionally GP based smoothed observations before and after outlier removal. The shaded region represents the 2σ area for the GP approach. Outliers are marked as circles.

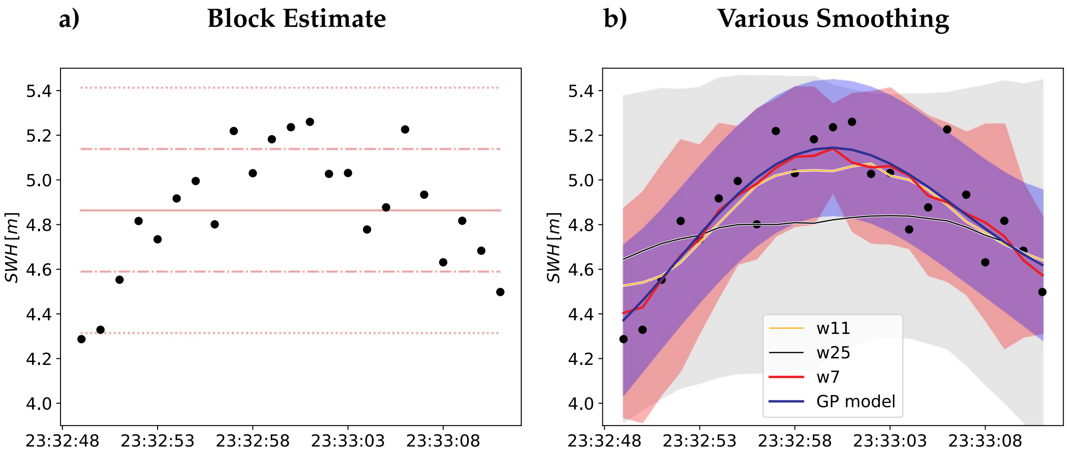

Fig. 5. a) S3a observations are displayed as black dots, the red line depicts the block mean, the stippled lines the 1σ and the dotted lines the 2σ deviation. b) S3a observations together with super observations displayed as lines: GP in blue, and the moving averages with window sizes 7 (red), 11 (orange), and 25 (black). The shaded regions illustrate the 2σ area for the GP approach (blue), moving average with window size 7 (red), and for window size 25 (gray).

3.2. Comparing super observations from both approaches

3.2.1. Example A: short time series with a local maximum

Fig. 5 b) shows super observations derived from the moving averaging approach with different block sizes together with the GP approach. If the displayed observations were the only available observations, multiple values close to the boundary could not exist based on this block size (illustrated in Fig. 4 a). The GP approach, however, can provide super observations for the entire given observation time period. To compare the different approaches for this case the moving average was allowed to borrow information from outside of the given block.

The choice of the averaging window is crucial for the resulting super observations where the similarity with the GP approach increases with decreasing block size (Fig. 5 b). However, the smaller the block size, the noisier are the super observations and the less representative is the sample size for computing the block average. The 2σ area reduces with decreasing block size but depicts more frequent abrupt changes in variance. These abrupt changes are challenging to interpret and probably unrealistic, as the moving average seems to be more certain close to the local maximum and will be highly influenced by single values.

3.2.2. Example B: longer time series

A satellite track with a realistic sample size (zoom in over the gray shaded area in Fig. 3) is shown in Fig. 6 a where the satellite track is directly compared with the wave model output. The average based super observations with a block size of 7 values represent the super observations which one would have chosen according to the commonly applied scale analysis. These moving average derived values follow closely the GP based super observations. However, for the GP based super observations no block size had to be chosen explicitly. Rather, a physically logical range can be given where the estimation of the length scale is a result of a maximum likelihood optimization based on the behavior of the observations. The resulting function carries the imprint of the information in the data through the optimal estimation of the trend and the three model parameters.

The two examples of block sizes illustrate what is intuitively anticipated, namely that the larger the block size the smoother the super observations. Choosing an inappropriate block size of 25 consecutive values irons out important features in the observed wave field. Choosing the block size of 7 values results in a noisier structure but more variability in the observational time series can be retained. The GP approach can model the observations with a very smooth function

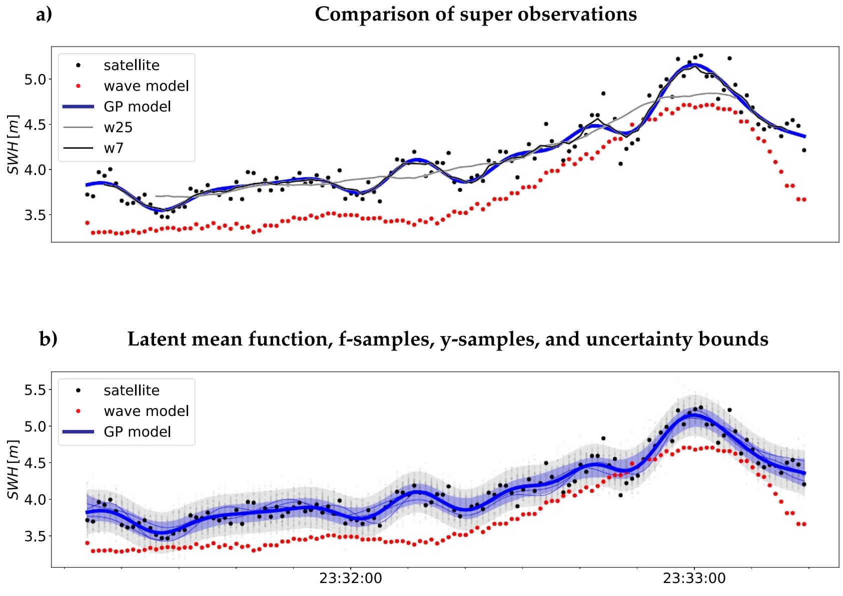

Fig. 6. SWH time series for S3a observations (block dots) and collocated wave model output (red dots). a) Comparison of super observations comprising the moving average derived super observations for window size 7 (black), 25 (gray), GP (blue). b) The latent mean function representing the GP derived super observations is depicted in blue. The gray dots represent realizations of observations based on the GP model as described in step 2 in Section 2.4.1. The gray shading describes the region between the 2σ uncertainty bounds of these realizations. The blue shading depicts the 2.5 and 97.5 percentile uncertainty bounds of the latent mean function derived from the samples in step 3 in Section 2.4.1.

while retaining the variability in SWH. Such a continuous behavior is expected for a SWH field where abrupt changes would only occur on very fine spatial scales under certain circumstances (e.g. modulation of the wave field by bathymetry, wave-current interactions, or sharp changes due to coastlines).

Another advantage of the GP approach is that it can provide information about the uncertainty of the observations and the super observations (latent mean function) as described in Section 2.4.1. Each observation and super observation can be seen in context with a desired number of realizations (Fig. 6 b). The scatter around the observations illustrates that each observation is just a random realization of what is likely to be the true sea state. This makes both the super observations and the observations probabilistic. The uncertainty can subsequently be propagated to other applications, such as validation purposes.

3.3. Implications for the validation of wave models

To illustrate the applicability and benefits of the GP approach for wave model validation, we compute validation statistics commonly included in validation reports. The validation statistics are defined in A and are computed based on the time series illustrated in Fig. 6.

When computing a validation statistic from average derived super observations, a single value describes the goodness of the wave model compared to an observational reference. This approach cannot provide an uncertainty estimate. This is problematic because the super observations used for the validation are based on semi-subjective decisions which naturally introduce uncertainty (different choices would lead to different super observations). This uncertainty should be propagated to the validation statistics in order to rigorously quantify the significance of their magnitude.

We present two attempts to assess the uncertainty in the computed validation statistics (Sections 2.4.2 and 2.4.1). The resulting distributions for each validation statistic can differ considerably between the approaches (Fig. 7). Significant discrepancies are visible for the scatter index and the correlation coefficient while results for the bias are very similar. The average based statistics overestimate the quality of the model compared to the raw data with non-parametric bootstrapping and the GP approach. For deviation-based statistics this means a lower value and for the correlation a higher value. This behavior results from smoothing the time series as illustrated in the following.

Proposed by Murphy (1988) the mean squared error (MSE) can be decomposed as follows:

$$MSE = (\overline{m} - \bar{o})^2 + \sigma_m^2 + \sigma_o^2 + 2\sigma_m\sigma_o r_{mo} \tag{6}$$

where $\overline{m}$ is the mean of the model time series, $\bar{o}$ is the mean of the observational time series, σ_m and σ_o represent the standard deviation of the model and observational time series (super observations or original), respectively. Pearson product-moment correlation between the model and observations is denoted by r. The validation statistic MSE and therefore also RMSE (A) depend directly on the variance within the compared time series and hence the smoother the time series the better the validation statistic. The, in our case inappropriate, window size of 25 therefore generates the best validation scores, while validating against the noisy raw S3a data systematically leads to the worst results.

The non-parametric bootstrap method applied to the raw observations can provide a distribution of the validation statistics with a substantial spread (Fig. 7, red histogram). The spread reflects the uncertainty based on the noise inherent in the observations. It is then possible to choose percentiles to express a probability. When comparing, however, the bootstrap results with the average and GP based results, a systematic mismatch becomes visible. As indicated above, this is due to smoothing the time series, but the question emerges whether the values being compared represent the same physical processes?

In case of using the non-averaged values for validation and performing a bootstrap, it is not taken into account that the model has a certain (effective) resolution. In this framework, the uncertainty and the associated histogram is not representative of the uncertainty of results from averaged or smoothed values. Essentially two different types of values are compared resulting in an overall underestimated model quality due to the noise in the observations. When only using average based results, the semi-subjective choice of the correct length scale is crucial to producing representative observational values which can be compared to model values on the wave model scale. Recalling that the chosen length scale is, besides being partly subjective, also static, it is likely that an inappropriate length scale is chosen in some cases. At the same time no uncertainty of the derived statistic is available. Note that one could possibly quantify such uncertainty using relatively advanced parametric bootstrapping (Davison and Hinkley, 1997). However, this is beyond the scope of this paper, where the aim is to illustrate the relative ease with which such uncertainties are obtained from the GP approach. The length scales chosen in this manuscript feature a tendency to overestimate the wave model quality. Likewise, an underestimation would be possible in some cases when choosing to small length scales.

These shortcomings from having to decide on a length scale can be mitigated by using the GP approach. If necessary the length scale for the GP approach can be constrained based on physical arguments. The optimization of the model parameters (signal variance, noise variance and length scale) are not subject to a subjective decision on a stationary value but are chosen objectively by maximizing the likelihood. Those parameters are allowed to vary from case to case and are learned each time depending directly on the characteristics of the observations. Moreover, the result is probabilistic and can provide a distribution expressing the uncertainty of the quality of the wave model.

4. Discussion

The chosen example featuring a strong gradient in the wave field due to the sheltering effect of Greenland might seem extreme. It is not, however, a single or unusual example. As mentioned before, a tip jet at the southern tip of Greenland is a feature of frequent occurrence (e.g. Våge et al. (2009)). Based on measurements from the NASA-JPL SeaWinds scatterometer on board the Quick Scatterometer (QuikSCAT) satellite, Moore and Renfrew (2005) stress that the tip jet around Greenland leads to surface wind speeds as high as 50 ms^{-1} with winds over 25 ms^{-1} occurring approximately 10% to 15% of the time. Those intense winds have a significant impact on the sea state for a considerable fraction of time when validating the wave model performance in this region. Especially due to the sheltering effect of land, sharp gradients are expected to occur other places in the world oceans.

When choosing a static block size for outlier detection, gradients and regime shifts in the wave field can violate the assumption of stationarity within the chosen block. A feature- and scale-aware length scale as presented here can take this into account as the information is accessible in the observations. Since each predicted value of the GP model is weighted by the behavior of the neighboring observations across the entire time series, the result is less sensitive to outliers while still reproducing the variability of the observational data to the desired degree. This is of particular importance as the effective wave model resolution advances to ever finer scales. The mismatch between the length scale of the satellite observations and the effective model resolution increases and could result in e.g. one or only a few satellite measurements per model grid point. Knowing, however, that the satellite value is a result of a stochastic process with considerable spread and abrupt changes, while the true sea state features smooth transitions, a limited number of observations might not be representative of the sea state. As model resolution continues to increase, this problem needs to be addressed and the GP approach is an attractive candidate.

An issue with smoothing by using a moving average, is that dependence between the errors of single observations are introduced (see Section 2.2). However, when planning to use the super observations for

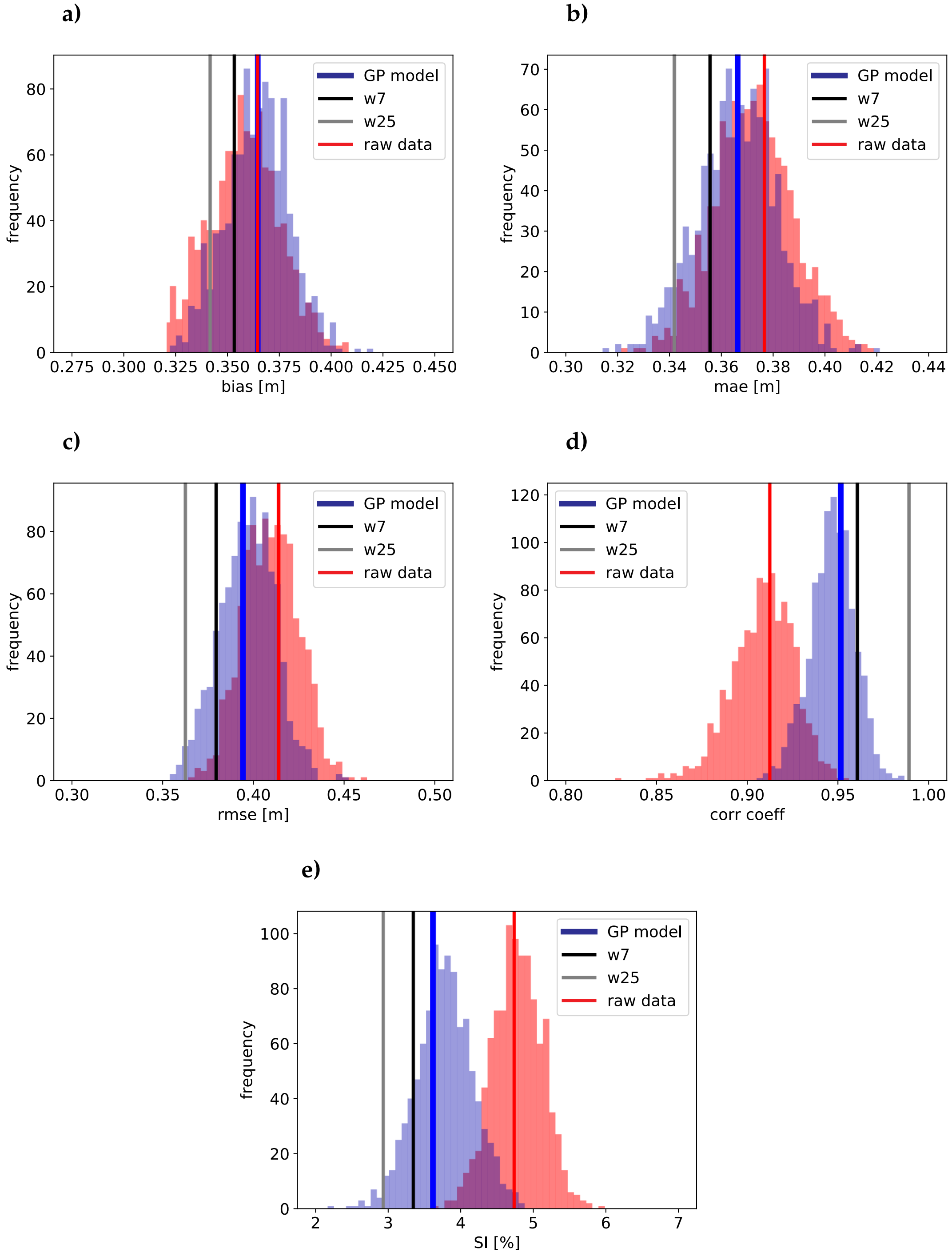

Fig. 7. Validation statistics computed with average based values with window sizes 7 and 25, non-parametric bootstrap results of raw observations, and results from the GP approach. Histograms are given in red for the non-parametric bootstrap results and in blue for the GP results. The violet shaded region indicates overlapping of the two histograms.

data assimilation into the wave model, independence of the errors is desired (Saleh Abdalla, 2018, personal communication). This is the reason for choosing non-overlapping blocks of values as opposed to a moving average. It is noteworthy that error dependency can be expected at every physical scale and that therefore even non-overlapping blocks cannot ensure independence of errors in general. The GP model as formulated here will create a smooth super observational time series without introducing additional error correlation as Gaussian noise is only added to the diagonals of the covariance matrix. The skill of the GP model to capture the auto-correlation inherent in the underlying process while leaving error dependency insignificant is displayed in Fig. B.8. The GP produced values can therefore be used for both wave model validation and data assimilation. Not having to divide into independent blocks further results in more super observations available for validation and assimilation.

The GP model introduced here is motivated by the physical properties of the wave field. However, as demonstrated in Rasmussen and Williams (2006) and Camps-Valls et al. (2016), there are many ways to adjust the kernel function to the problem at hand. Our work aims mainly at introducing the GP approach to the field of wave model validation using super-observations and presenting inherent advantages. For this reason, we chose a comparably simple model formulation which serves the understandability of what is going on in the GP-model. To optimize the performance of the space-time evolution one could explore e.g. the impact of using different kernel functions as well as customize a desired behavior by combing the available kernel functions to imitate the presumed underlying process. A full Bayesian inference (Rasmussen and Williams, 2006) could also be pursued where the kernel function parameters are again parameterized with hyperprior distributions. As visible in the Fig. B.8, despite the overall good performance of our GP-model there might be room for tuning, e.g. regarding the auto-correlation of the y-samples within the first lags and the wiggles in the residuals. However, it is important to note that neither of these issues, wiggles nor the first lags, are of statistical significance. In future studies, especially the impact of adding heteroscedasticity to the GP model could be investigated.

5. Summary and conclusion

We propose a highly flexible approach to compute super observations and detect outliers utilizing a Gaussian Process. We subsequently validate wave model results where our approach allows us to retrieve a probabilistic estimate of the validation statistics. There is no need to exactly specify the length scale for the computation of super observations and outlier detection since our super observations and uncertainty bounds are maximum likelihood estimates. The data dependence and automated way to fit is scale-sensitive and assures flexibility. These are characteristics that are important for the validation of high resolution model simulations. If needed, our approach would even allow the prediction of super observations for a model grid cell along the satellite track, interesting e.g. for the validation of very high resolution simulations or for imputation of missing values. The characteristics of the here proposed formulation ensures that no significant additional error correlation is being introduced which is beneficial for using the super observations for data assimilation.

Gaussian Processes are getting increasing attention as state-of-the-art tools for regression problems. In the field of earth observation data analysis their advantages are discovered and exploited across scientific disciplines (Camps-Valls et al., 2016). Although our work focuses on the computation of super observations from satellite tracks for wave model validation, the here formulated GP model can be used just as well for measurements from other platforms e.g. buoys or other variables like e.g. wind speed.

Acknowledgement

The authors gratefully acknowledge the funding support from the Copernicus Marine Environmental and Monitoring Service under Mercator Ocan contract no. 60-CMEMS MFC ARCTIC. Øyvind Breivik acknowledges the support from the CMEMS Service Evolution project WaveFlow. We would also like to thank Saleh Abdalla (ECMWF) for helpful comments and suggestions at an early stage of this study.

Appendix A. Equations for the validation statistics

Variable names and the statistics Bias, MAE, and RMSE follow the convention from WMO (2010). The computation of the scatter index is adopted from the ECMWF.[3]

- x_f is the forecast values of the chosen parameter
- x_v is the value to evaluate against (“ground truth”)
- x_c is the climatological value of the chosen parameter
- n is the number of values to be used for verification

Bias or Mean Error:

$$\text{Bias} = \text{ME} = \frac{1}{n}\sum_{i=1}^{n}(x_f - x_v)_i \tag{A.1}$$

Mean Absolute Error:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|x_f - x_v|_i \tag{A.2}$$

Root Mean Square Error:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_f - x_v)_i^2} \tag{A.3}$$

Pearson product-moment correlation coefficient:

$$\text{r}_{\text{f,v}} = \frac{cov(x_f, x_v)}{\sigma_f \cdot \sigma_v} \tag{A.4}$$

Scatter index:

$$\text{SI} = \frac{std(x_f - x_v)}{mean(x_v)}\cdot 100 \tag{A.5}$$

Appendix B. Auto-correlation

This section illustrates the empirical auto-correlation function (acf) of the observational time series compared to the acf of artificially generated observations (y_i) produced by our GP-model. Consistent with the rest of this study, we produced 1000 artificial observational time series from a parametric bootstrap. The residuals were computed by subtracting original observations from all 1000 time series.

[3] https://www.ecmwfr.int/en/newsletter/150/meteorology/twenty-one-years-wave-forecast-verification.

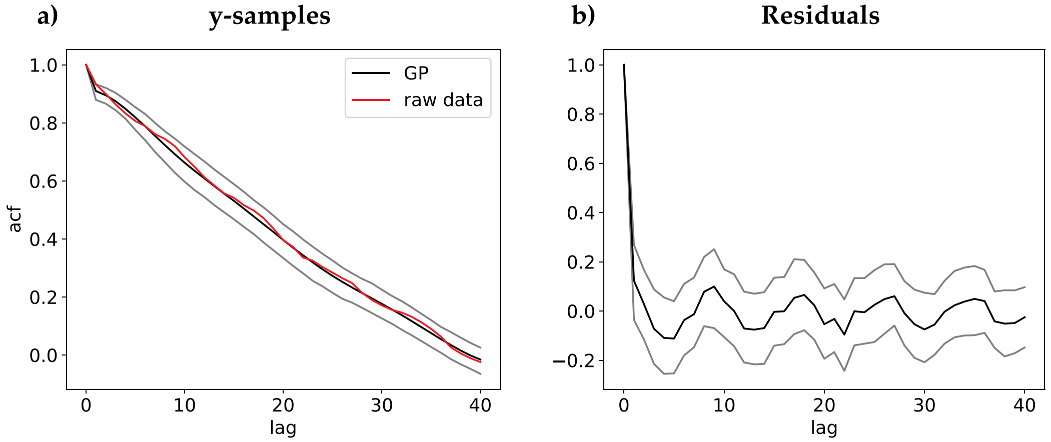

Fig. B.8. a) shows the mean of the empirical auto-correlation functions (acf) of 1000 y-samples generated with our GP model (black) with 95% uncertainty bounds (gray) from parametric bootstrapping. The red line is the acf from the unprocessed measurements depicting a very similar behavior. b) shows the acf of the residuals where no significant auto-correlation is present.

References

Abdalla, Saleh, 2018. S3-A Wind & Wave Cyclic Performance Report.

Abdalla, S., Janssen, P.A.E.M., Bidlot, J.-R., 2011. Altimeter near real time wind and wave products: random error estimation. Mar. Geod. 34 (3–4), 393–406. https://doi.org/10.1080/01490419.2011.585113.

Abdalla, S., Isaksen, L., Janssen, P.A.E.M., Wedi, N., 2013. Effective spectral resolution of ECMWF atmospheric forecast models. ECMWF Newslett. 137, 19–22.

Camps-Valls, G., Verrelst, J., Munoz-Mari, J., Laparra, V., Mateo-Jimenez, F., Gomez-Dans, J., 2016. A survey on Gaussian processes for earth-observation data analysis: a comprehensive investigation. IEEE Geosci. Remote Sens. Mag. 4 (2), 58–78. https://doi.org/10.1109/MGRS.2015.2510084.

Davison, A.C., Hinkley, D.V., 1997. Bootstrap methods and their application. In: Cambridge Series in Statistical and Probabilistic Mathematics.

Doyle, J.D., Shapiro, M.A., 1999. Flow response to large-scale topography: the Greenland tip jet. Tellus A 51 (5), 728–748. https://doi.org/10.1034/j.1600–0870.1996.00014.x.

Efron, B., 1979. Bootstrap methods: another look at the jackknife. Ann. Stat. 7 (1), 1–26. https://doi.org/10.1214/aos/1176344552.

Holthuijsen, L.H., 2010. Waves in Oceanic and Coastal Waters. Cambridge University Press.

Komen, G.J., Cavaleri, L., Donelan, M., Hasselmann, K., Hasselmann, S., Janssen, P., 1994. Dynamics and Modelling of Ocean Waves. University Press, Cambridge, UK, Cambridge0521470471.

Liu, Q., Babanin, A.V., Zieger, S., Young, I.R., Guan, C., 2016. Wind and wave climate in the Arctic Ocean as observed by altimeters. J. Clim. 29 (22), 7957–7975. https://doi.org/10.1175/JCLI–D–16–0219.1.

Moore, G., Renfrew, I., 2005. Tip jets and barrier winds: a QuikSCAT climatology of high wind speed events around Greenland. J. Clim. 18 (18), 3713–3725. https://doi.org/10.1175/JCLI3455.1.

Murphy, A.H., 1988. Skill scores based on the mean square error and their relationships to the correlation coefficient. Mon. Weather Rev. 116 (12), 2417–2424. https://doi.org/10.1175/1520–0493(1988)116 < 2417:SSBOTM > 2.0.CO;2.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.

Pickart, R.S., Spall, M.A., Ribergaard, M.H., Moore, G., Milliff, R.F., 2003. Deep convection in the Irminger Sea forced by the Greenland tip jet. Nature 424 (6945), 152.

Rasmussen, C.E., Williams, C.K., 2006. Gaussian Processes for Machine Learning. 2006. 38. The MIT Press, Cambridge, MA, USA, pp. 715–719.

Stopa, J.E., Ardhuin, F., Girard-Ardhuin, F., 2016. Wave climate in the Arctic 1992–2014: seasonality and trends. Cryosphere 10 (4). https://doi.org/10.5194/tc–10–1605–2016.

Våge, K., Spengler, T., Davies, H.C., Pickart, R.S., 2009. Multi-event analysis of the westerly Greenland tip jet based upon 45 winters in ERA-40. Q. J. R. Meteorol. Soc. 135 (645), 1999–2011. https://doi.org/10.1002/qj.488.

WMO, 2010. Manual on the global data-processing and forecasting system. Volume 1-global aspects. WMO 485.

Young, I., 1999. Seasonal variability of the global ocean wind and wave climate. Int. J. Climatol. 19 (9), 931–950. https://doi.org/10.1002/(SICI)1097–0088(199907)19:9 < 931::AID–JOC412 > 3.0.CO;2–O.

Zieger, S., Vinoth, J., Young, I., 2009. Joint calibration of multiplatform altimeter measurements of wind speed and wave height over the past 20 years. J. Atmos. Ocean. Technol. 26 (12), 2549–2564. https://doi.org/10.1175/2009JTECHA1303.1.