Rapid response data-driven predictions for storm surge around New Zealand.

Tausia J[1], Delaux S[2], Camus P[1], Rueda A[1], Mendez F[1], Bryan K[3], Perez J[2], Costa C. G. R.[2], Cofino A[4], and Zingfogel R[5]

[1]Geomatic and Oceanographic Engineering Group, University of Cantabria, Spain
[2]Meteorological Service of New Zealand, New Zealand
[3]University of Waikato, Hamilton Waikato, New Zealand
[4]Department of applied mathematics and computer science, University of Cantabria, Spain
[5]Calypso science, New Plymouth, New Zealand

Correspondence: Javier Tausia (tausiaj@unican.es)

Abstract. In conjunction with tides, storm surge is one major driver of coastal flooding associated with storm events. Because local inundation is strongly modulated by the local shape of the coastline and the bathymetric slope, accurate storm surge predictions using traditional numerical models require the use of very fine grids and is hence resource intensive. Therefore, the performance of a live prediction system based on such methods will likely be subject to a trade-off between prediction accuracy, prediction speed and cost.

In this study, we explore the use of 3 data driven methods as an alternative to numerical methods to reconstruct the 6 hourly, 12 hourly and daily storm surge maximum levels along different locations in the coast of New Zealand. We first explore different atmospheric predictors with these 3 statistical models, to find the best possible predictor, and then, we reconstruct the storm surge daily maxima with the different statistical models along the entire coast, based on this predictor.

The code developed and used as part of this study is public and available in a GitHub repository to facilitate easy replication of the study.

KEYWORDS - Data-driven models, storm surge, atmospheric predictor, rapid spatial reconstructions

1 Introduction

With over 15,000 km of coastline and around 150,000 people living in low-lying coastal areas, coastal inundation is a major hazard to New Zealand (NZ). The cost to defend the associated buildings, infrastructure and assets is of the order of $10 billion (Ministry for the Environment (NZ), 2017, Preparing for Coastal Change, Publication number: ME 1335, 36pp.). With global sea level rise and the increase in the intensity and frequency of extreme weather events expected with climate change, the threat posed by coastal flooding is only awaited to become greater.

Storm surge is the rise of water level generated by wind and atmospheric pressure changes associated with tropical or extra-tropical (mid-latitude) storms, over and above the astronomical tide (AT), and the long-term signals as the monthly mean sea level accounting for the seasonal and inter-annual variability (Cid et al., 2017). Storm surge is one of the most critical

components of coastal flooding and its magnitude has a large spatial variability (Bell and Goring, 1996). Flooding associated with storm surges is one of the most common natural hazards for coastal areas worldwide (Bell et al., 2000).

In New Zealand, AT accounts for 96% of the coastal energy, while the over-elevation associated to barometric pressure and wind effects, the effect of waves, see Stephens et al. (2011), and the longer-term seasonal and inter-annual fluctuations account for the remaining 4% (Bell et al., 2000; Goring and Bell, 1996). Although storm surge around New Zealand, reaching just 0.8 m maxima above mean sea level, see Heath (1979), is much lower than storm surge experienced in equatorial regions and high latitudes, it can still cause coastal flooding and exacerbate coastal erosion (Bell et al., 2000). For example, a flooding event that occurred in 1995 in the Thames Region, when peak storm surge overlapped with high AT, entailed damages worth around 3–4 million dollars, see Bell et al. (2000). In addition, during the spring and summer of 2017 and 2018, several large storms including ex-tropical cyclones Fehi, Gita, and Hola struck NZ, most of them coinciding with high perigean-spring tides, causing flooding to homes and damaging infrastructure. Other notable historical coastal flooding events in NZ occurred in January 2011, during cyclone Gisele in 1968, see de Lange and Gibb (2000), May 1938 in the Hauraki Plains (Stephens et al., 2020) and during the great cyclone of 1936 (Brenstrum, 2000), but the spatial effects of these historical storms are not well recorded since not many sea-level gauges were in operation at those times, see Stephens et al. (2019a).

Sea level forecasts usually use computationally expensive numerical models, which require running the model given the predicted atmospheric conditions every time a prediction is to be made. Local inundation is strongly modulated by the local shape of the coastline and the bathymetric slope, and so accurate storm surge predictions by the mean of traditional numerical models require the use of very fine grids and is hence resource intensive. This means that the performance of a live prediction system based on such methods will likely be subject to a trade-off between prediction accuracy, prediction speed and cost (Wang et al., 2009). In this sense, this study seeks to find the most efficient atmospheric predictor for reconstructing storm surge maximum levels in New Zealand, using 3 different statistical methods, and with the aim of producing a reliability similar to that of numerical models, see Siek (2019), but at a fraction of their computational effort.

To date, there are many studies that reconstruct, given an atmospheric predictor, variables such as storm surge (Cid et al., 2018, 2017), significant wave height (Camus et al., 2014) or even both (Rueda et al., 2019), using statistical linear models, but none of these studies explain why the used predictor is better than any other possible combination of atmospheric variables. These studies use linear models to reconstruct the variables of interest, while other studies such as Bruneau et al. (2020) use neural networks for the same purpose (using pre-defined predictors too). Bruneau et al. (2020) obtain promising results for the sea level at a large number of locations over the world, because they capture the non-linear relationships between the predictor and the predictand. Cagigal et al. (2020) reconstruct the storm surge over NZ obtaining decent results for different locations around the islands, but again, a single predictor is used.

There are recent studies that compare the predictive capabilities of several data driven methods. Tadesse and Wahl (2021) show how different predictors can reconstruct the storm surge around the world, giving a set of potential statistical models, and Tiggeloven et al. (2021) also evaluate predictor capabilities using neural networks as the main model, obtaining better results as the predictor gets bigger in space and includes more variables such as the wind, or non linear components of these used

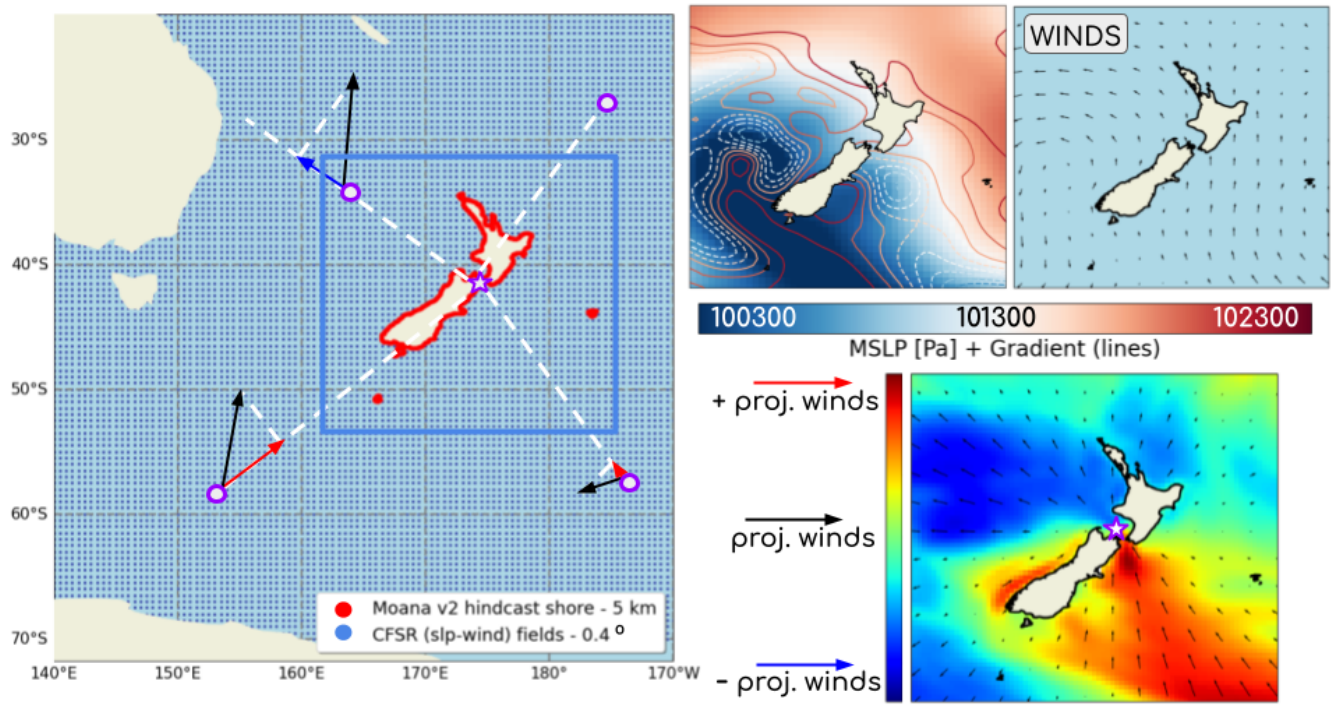

Figure 1. The atmospheric reanalysis is represented by blue dots (blue square represents the maximum spatial domain considered in the experiments), while the Moana v2 hindcast model output, which correspond to the storm surge hindcast, is shown in red. The two plots in the top-right represent the sea-level-pressure fields (SLP) and the corresponding gradients (left), and the u10 and v10 components of the wind are also shown (right). Finally, the "projected winds" are shown. All the figures in the right correspond to the same time in the atmospheric reanalysis. See text for descriptions of projected wind vectors.

variables (u^2, v^2...). Nevertheless, these studies do not explain why the differences in the results given the different predictors and models might appear, lacking a comprehensive framework on which to base such an assessment.

This study tests 108 different atmospheric predictors in order to get an understanding of the contribution of the different variables to the predictor set (see Figure 1). The parameters are: variables (sea-level-pressure; SLP, gradients of the sea-level-pressure fields and projected winds), time lapse (i.e., how many past time frames are used for each single prediction), time resampling (6-hourly, 12-hourly and daily), and spatial extent (local 3·3º, local 5·5º and regional covering the whole spatial area of New Zealand). Additionally, 3 different statistical models, including linear regression, k-NN regression and gradient boosting regression, with their own extra hyperparameters, are trialled too.

Model performance is evaluated with the Modified Kling-Gupta Efficiency, see Kling et al. (2012), as it incorporates 3 sub-metrics; correlation, bias term and variability term, each characterising an important aspect of the prediction performance. This detailed predictor analysis is performed in 29 different locations around New Zealand, where nearby observational data is also

available. Once the best atmospheric predictor is identified the reconstruction is extended to the whole of the New Zealand coastline and the results are contrasted for the 3 statistical models.

This article is structured in 6 sections. In Section 2, the databases used are described, then, in Section 3, the methodology followed is explained in detail. In Section 4, results for all the predictors, models and locations are shown. In Section 5, everything is summarized in the discussion, where the final thoughts for the results obtained are shown. Finally, future tasks and conclusions are captured in Section 6. At the end of the paper, in the Appendix, the detailed explanation for all the statistical models used can be found.

2 Databases description

Model predictors data were sourced from a global atmospheric reanalysis. Predictands were obtained from a high resolution regional hydrodynamic hindcast for New Zealand waters. In Figure 1, the spatial domains of both hindcasts can be seen, where variables are also shown. In addition, observations acquired from 29 tidal gauges spread around the coast of New Zealand (see Fig.2) were used to validate the sea level data from the hindcast.

2.1 Atmospheric data

For the atmospheric data, we used a global reanalysis developed by NCEP (National Centers for Environmental Prediction) in the configuration of CFSR (Climate Forecast System Reanalysis), Saha et al. (2010), which extends from 1979 to 2011, and CFSRv2, Saha et al. (2011), which extends from 2011 to present. Variables in both datasets are similar, and they are both global reanalysis for atmospheric conditions. From the different variables, we have chosen sea-level-pressure and winds as the main predictors affecting the storm surge. SLP data are at the same resolution of 0.4º in both datasets and hence were kept in their original configuration. However, the eastward and northward components of the wind, whose original resolution is 0.3º in CFSR and 0.2º in CFSRv2, were interpolated linearly over the same grid as the SLP.

These wind components are not directly used, but they are projected to the location where the reconstruction is made. As illustrated in Fig.1, the wind vectors are projected over the line that joins each point in the atmospheric gridded domain to the location where the storm surge is predicted. Then, if the wind points to a desired location at time t, this wind will highly contribute to the storm surge signal (red arrows). On the other hand, winds blowing in the direction opposite to the desired location will not contribute to the storm surge (blue arrows). Finally, land location is also taken into account, so winds directly blowing towards a certain location but from land side are discarded.

2.2 Storm surge data

The storm surge datasets are separated into hindcast and observational data. The hindcast is preferred here for its spatial and temporal robustness, as we want to reconstruct the storm surge maximum levels all over the New Zealand coast, so it is used to calibrate the statistical models (other studies use tidal gauges to calibrate the models, see Cagigal et al. (2020) and Tiggeloven et al. (2021), for example). The observational data, which correspond to tidal gauges, is used to validate this hindcast. As shown

in Figure 2, comparisons exhibit a very good correlation between the outputs of the numerical model and the observational data, and thus this storm surge hindcast can be used to calibrate the statistical models (metrics such as the RMSE, the Pearson or Spearman correlations and many other statistics can be found in the GitHub repository for all the available nodes).

The processing of the total sea level series for both the hindcast and the tidal gauges data was done using the open-source toolbox Toto (https://github.com/calypso-science/Toto). The linear trend was first removed from the time series. Tidal analysis was then carried out using the algorithms implemented in the Python version of the UTide software (Codiga, 2011) and the astronomical tide estimates were used to fill any missing gaps in the tidal gauge data. The monthly mean sea level variation was then removed from the time series using a Lanczos filter. Finally, the storm surge signal was extracted using a Lanczos lowpass filter, see Thomson and Emery (2014), with a cut-off period of 30 hours. Considering that the inertial period around New Zealand latitudes varies from 16h to 22h approximately, the 30 hours cut-off period allowed for both tidal and inertial oscillations to be removed from the total water level, isolating the storm surge signal.

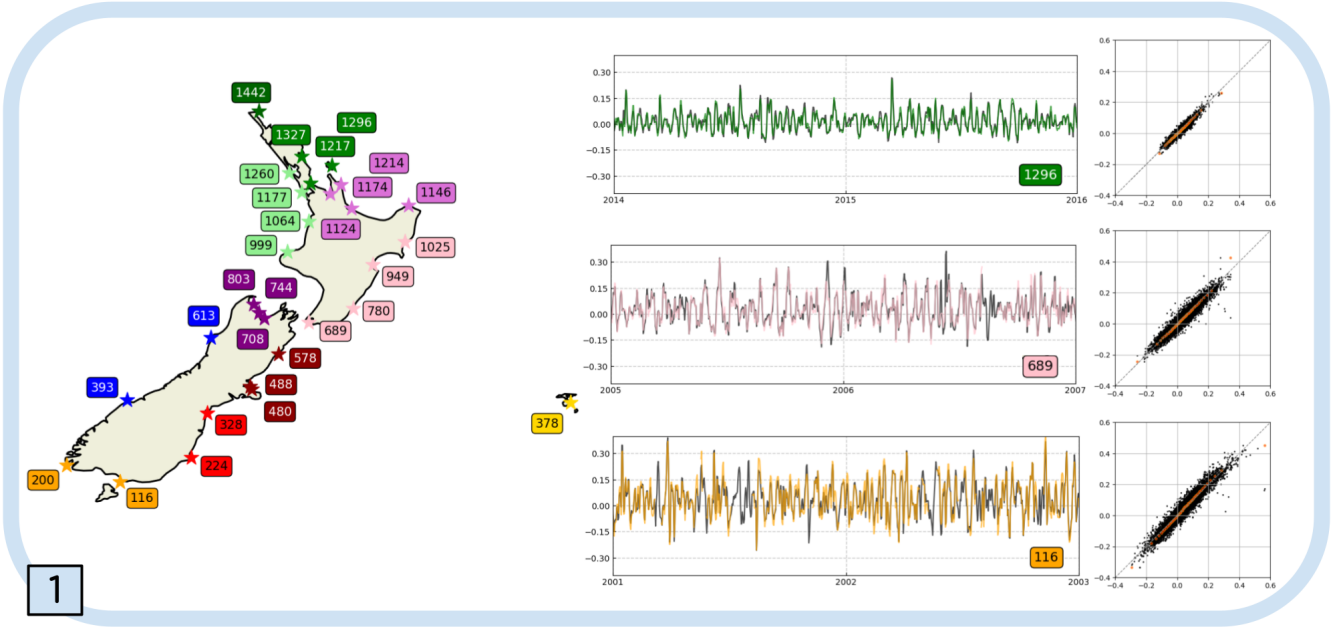

Figure 2. Location of tidal gauges along the NZ coast and its closest numerical model nodes (left) and validation of Moana vs tidal gauge in three locations (right). Hindcast is represented in black in the time series plots.

2.2.1 Moana v2 hindcast

The numerical model storm surge data used to train the models along the New Zealand coast was obtained by postprocessing of the sea surface height fields from version 2 of the Moana Backbone Model, see de Souza (2022). The Moana Backbone Model is a 25-year regional hydrodynamic hindcast model of New Zealand waters released in 2020 by the New Zealand MetService. The hindcast was produced using the Regional Ocean Modeling System (ROMS), version 3.9, which is a free-

surface, terrain-following, hydrostatic numerical model that solves the 3D Reynolds-averaged Navier-Stokes equations using Boussinesq approximation, see Haidvogel et al. (2008). The hindcast horizontal resolution is 5 km over the whole domain with 50 levels in the vertical. ROMS was forced with atmospheric conditions from the Climate Forecast System Reanalysis (CFSR) versions 1, Saha et al. (2010), and 2, Saha et al. (2011). The open boundaries were forced with currents, sea level, temperature and salinity, from the Copernicus Global Ocean Physics Reanalysis (GLORYS) version 12v1 and spectral tidal forcing from the OSU Tidal Inversion software (OTIS) version 7.1.

2.2.2 Observational data

We gathered data from 29 tidal gauges located around NZ as shown in Fig.2 and were used to validate the numerical model. We also used the location of the tidal gauges to select the closest hindcast nodes for which the initial predictor experiments were performed. This was motivated by the fact that the different sub-shores where these tidal gauges are located exhibit different storm surge behaviours, as the complexity and varying orientation of New Zealand's coastline mean that there can be strong local differences in storm surge signals, and the TGs are well spread around NZ. Also, this leaves us the option of validating results against nearby measured data.

3 Methodology

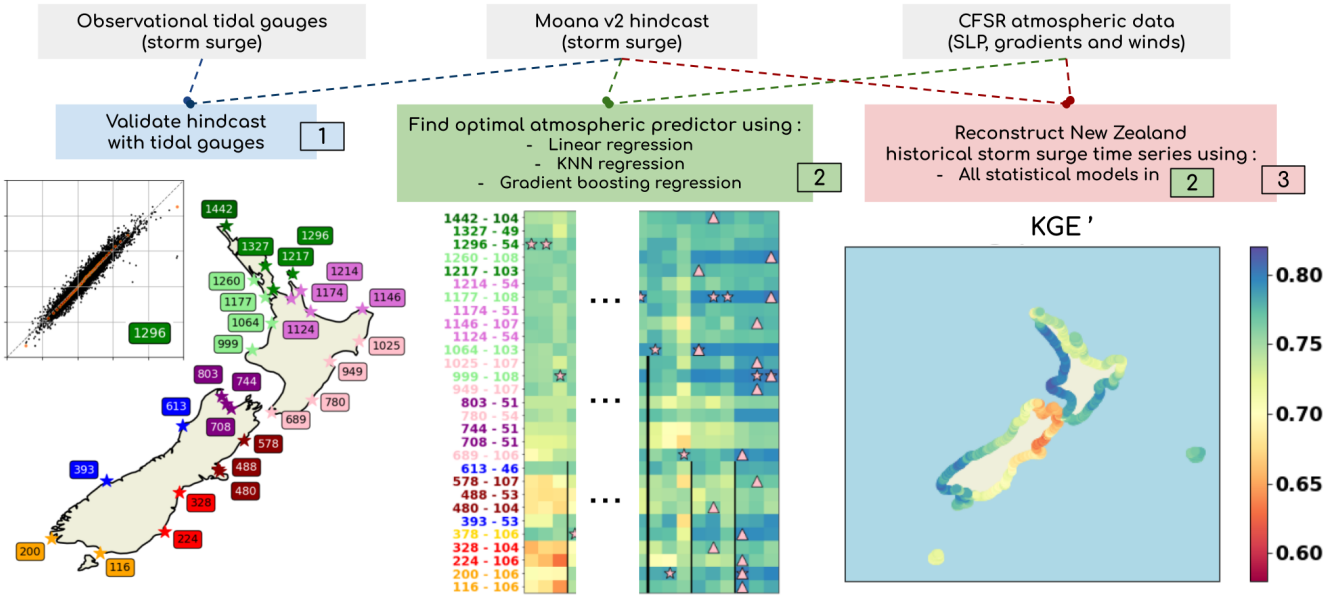

Figure 3. A schematic of the analysis workflow. 1) Storm surge hindcast validation with the tidal gauges. 2) The best performing atmospheric predictor is found. 3) With this predictor, the storm surge is reconstructed all over the NZ coast.

The process that was followed is divided in three parts shown in Figure 3. We first validate the numerical model outputs, then find the best possible atmospheric predictor, and finally reconstruct the storm surge maximum levels along the entire coast of New Zealand. These main parts are further subdivided below.

3.1 Moana v2 hindcast validation

All the available tidal gauges are used to validate the storm surge hindcast, obtaining very good results so allowing this hindcast to be used to calibrate the statistical models. Regarding the goodness of the validation plots in Fig.2, all the nodes have a Pearson correlation bigger than 0.92. Additional validation statistics can be found in the GitHub repository of the project.

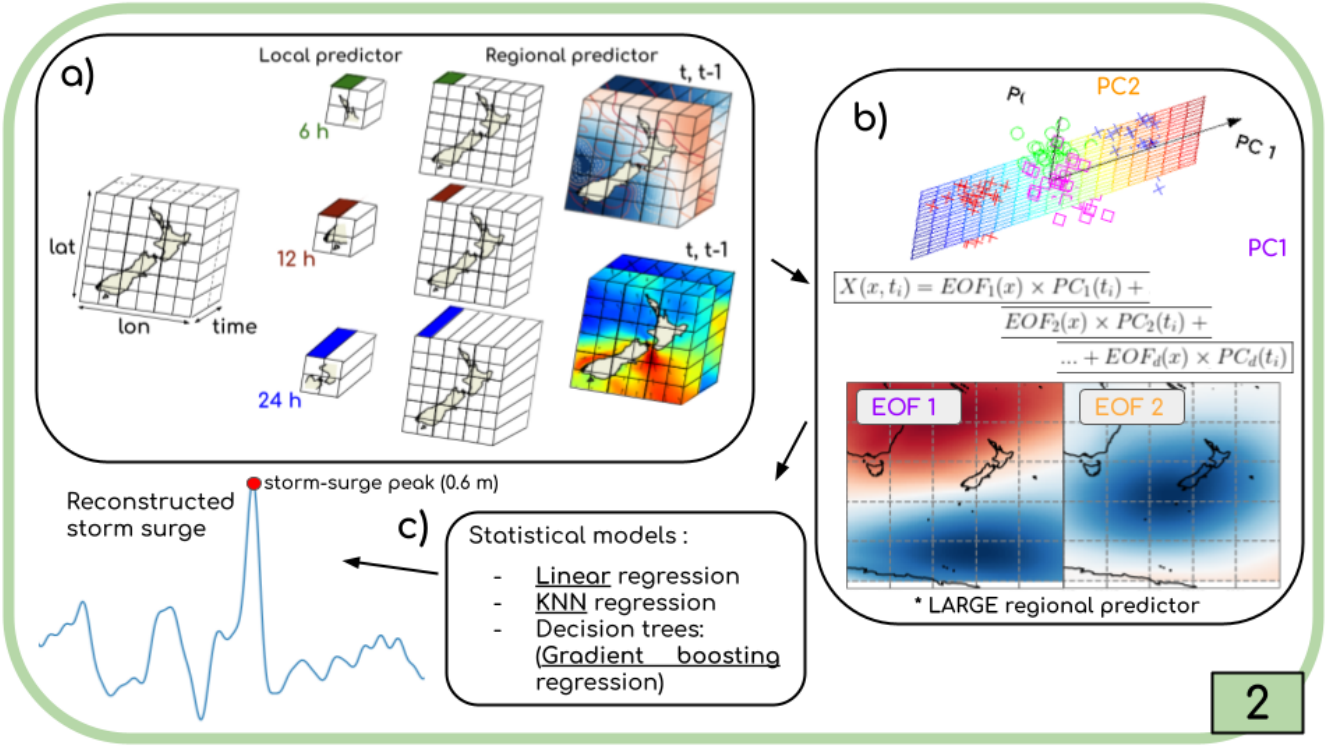

Figure 4. a) This is the structure that might take the atmospheric predictors, b) which are then passed through PCA, extracting the first PCs that explain a very high percentage of the variance in data. c) Finally, the storm surge is reconstructed using different statistical methods in a regression framework.

3.2 Optimal atmospheric predictor analysis

Once the storm surge hindcast data is validated, we identify the optimal atmospheric predictor from a total of more than 9000 experiments, which include any possible combination of atmospheric predictor, location and statistical model. Like that, having 108 atmospheric predictors (see Table 1), 29 locations to study and 3 statistical models to evaluate (linear, k-NN and gradient

boosting regression), this makes a total of 9396 experiments, but there are some of the models that have hyperparameters to tune, so in those cases, the number of experiments to run is greater.

The same workflow is followed by all linear methods as depicted on Figure 4 and involves: a) assembling the predictor, b) a dimensionality reduction step using Principal Components Analysis (PCA) (Gutiérrez et al., 2004; Wilks, 2005) and c) training the model against the PCA-projected predictor. The three steps are explained in detail below as they summarize the way the optimal atmospheric predictor is found.

3.2.1 Predictor building and PCA (a and b)

The search for the optimal atmospheric predictor is one of the main goals of this work, and hence is why all relevant combinations have been considered. This combinatorial cloud covers all the possible cases summarised in Table 1, where the parameters to be taken into account when constructing the predictor, and the values these parameters may take, are shown. The sea-level-pressure fields are always used, then we can add the gradients and the projected winds. For all these variables, past time frames can be used too (it is called time lapse in this study), so the information of previous times to the prediction are used. Regarding the spatial extent of the predictor, 3 different regions are trialled, two local squared regions of 3·3º and 5·5º centered in the location of interest and a bigger region that encompass the whole of New Zealand (blue square in Fig.1). Finally, different time re-samplings are used, so trying to reconstruct the storm surge maxima for different temporal resolutions, given the same shape of the predictors and models.

Table 1. Parameters used to construct the predictors are shown. Notice the SLP fields are always used, but all the other parameters might change.

4x Data sources			3x Time lapse	3x Time resample	3x Region
sea-level-pressure fields (SLP)	gradient fields calculated from the SLP variations	projected winds (calculated from u10 and v10)	whether to add previous time steps to reconstruct	time resolution to resample the data to reconstruct	spatial region to consider around the location of interest
predictor might use: SLP SLP + gradients SLP + projected winds **SLP + gradients + projected winds**			1 (just time t)	6 H	local - 3º x 3º
			2 (t and t-1)	12 H	**local - 5º x 5º**
			3 (t, t-1 and t-2)	**1 D**	Regional (160,185,-52,-30)

* this is the best performing predictor

As we have memory limitations, we had to work with a maximum time lapse of 3, so using the time of the prediction, and two past time frames of information, as there were cases when the existing machines could not resolve PCA. In this line, the time resolution to which data is resampled, so 24H, 12H and 6H, affect the amount of information used in the predictor; higher resolutions create less robust predictors, and lower resolutions allow for predictors with information of more "natural" days. For example, if we are working with 6 hourly resampled data, and we use 3 time frames to predict, we are using just 18 hours of information.

As a summary, the atmospheric predictor is built as a 2D matrix given a set of parameters that define its entries (rows due to the time resampling to 6H, 12H or 24H) and variables (number of columns depending on the spatial and temporal scales, plus the variables used).

After the initial predictor matrix is assembled for each experiment by concatenating the raw predictor data from the atmospheric reanalysis and in order to reduce the number of features fed to the statistical models, a dimensionality reduction step is applied to the predictor matrix. We use PC Analysis (see the Appendix for more detailed information) and keep the leading components ensuring that 98% of the variance is explained.

3.2.2 Statistical models (c)

The projected (PCA transformed) predictor is calculated, and with the N first principal components, the 3 previously mentioned statistical models can be trained given the storm surge maxima over the historical time range. In this study, we use the 70% of the data for training, and the rest 30% is used to validate the results. These validation metrics are the ones shown in all the existing plots.

We use linear regression, k-NN regression and gradient boosting regression algorithms to obtain the best possible atmospheric predictor from all the candidates. A brief explanation of these models is given below, while a more detailed one can be found in the Appendix.

1. **Linear regression**: The method fits the hyperplane (a hyper-dimensional plane with dimensions equal to the number of PCs used, N) to the data so the squared errors between the predicted and the real values are minimized. The best found linear regression parameters are given by $\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$.

2. **k-NN regression**: Given a dataset, each time a new prediction is made, the model predicts the mean of the target values of the k-nearest neighbors based on the Euclidean distance, see Altman (1992). In this study, we experiment with number of neighbors varying from 1 to 50, so 50 models are calibrated for each combination of predictor and location.

3. **Gradient boosting regression**: The method combines the ability of weak decision trees to obtain robust predictions of a target variable, optimizing the tree's individual ability by minimizing a loss function, calculating its gradient at each step, see Friedman (2000). In this case, we change both the maximum depth of the tree, testing 6, 12 and 18 final nodes, and the minimum percentage of the data available at each split, testing 2, 6 and 10% (trying to see how overfitting might affect the results). Then, 9 models are trained for each combination of predictor and location.

3.3 Storm surge spatial reconstruction in New Zealand

Finally, and given the best atmospheric predictor previously found, the storm surge is reconstructed over the whole coastal domain of the hindcast, using the above mentioned statistical models. The procedure involves training separate models for each of the coastal points following the workflow depicted in Figure 4, but using just the best atmospheric predictor. In this part, we use again the 70% of the data to train the models (which correspond to almost 17 years of data), and the remaining 30% is used to validate results. All metrics shown in this study are validation metrics.

All the hyperparameters for the different statistical models explained before are used here too, and results for all the coastal points and for this best predictor can be found in Figure 8.

3.4 Models evaluation

The Modified Kling-Gupta Efficiency statistic, see Kling et al. (2012), was chosen as the main metric to evaluate the models results and is defined as:

$$\text{KGE'} = 1 - \sqrt{(r-1)^2 + (\gamma-1)^2 + (\beta-1)^2} \tag{1}$$

$$\text{where } r = \frac{\sum_{i=1}^{p}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{p}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{p}(y_i - \bar{y})^2}}, \gamma = \frac{CV_{rec}}{CV_{obs}} = \frac{\sigma_{rec}/\mu_{rec}}{\sigma_{obs}/\mu_{obs}} \text{ and } \beta = \frac{\mu_{rec}}{\mu_{obs}} \tag{2}$$

where KGE' is the Modified Kling-Gupta Efficiency statistic (dimensionless, Kling et al. (2012)), r is the correlation coefficient between reconstructed and observed storm surge data (dimensionless), β is the bias ratio (dimensionless), γ is the variability ratio (dimensionless), μ is the mean, σ is the standard deviation, CV is the coefficient of variation (dimensionless), and the indices rec and obs represent reconstructed and observed storm surge values, respectively. KGE', r, β and γ have their optimum at unity. For the variability ratio γ we used CV_{rec}/CV_{obs} instead of $\sigma_{rec}/\sigma_{obs}$, which was proposed in the original version of the KGE-statistic (Kling-Gupta Efficiency, Gupta et al. (2009)). This ensures that the bias and variability ratios are not cross-correlated, which otherwise may occur when e.g. the atmospheric inputs are biased.

KGE' is particularly suitable here as it accounts for three different important aspects of the reconstructed time series by including a correlation, a bias and a variability term, and then giving a perspective on how well the distribution of the storm surge time-series is reproduced. For a full discussion of the KGE'-statistic and its advantages over the Nash–Sutcliffe efficiency see Nash and Sutcliffe (1970) (or with the mean squared error see Gupta et al. (2009)), as these are the most used metrics for evaluating hydrological model performance, where the behaviour of the models in the extremes is very important too.

4 Results

Results are analyzed in two different sections (sections 2 and 3 in Fig.3), as they were previously defined in the methodology. The results of the experiments aiming at identifying the best predictor are first analyzed, and then, the results of their extension to the whole coast, with the optimal atmospheric predictor, are presented.

4.1 Experiments results

The results of the experiments run for the different models and predictors at all selected locations are presented in Figure 5, where we can observe the relationship existing between the amount of data that is used as input and the best performing statistical model.

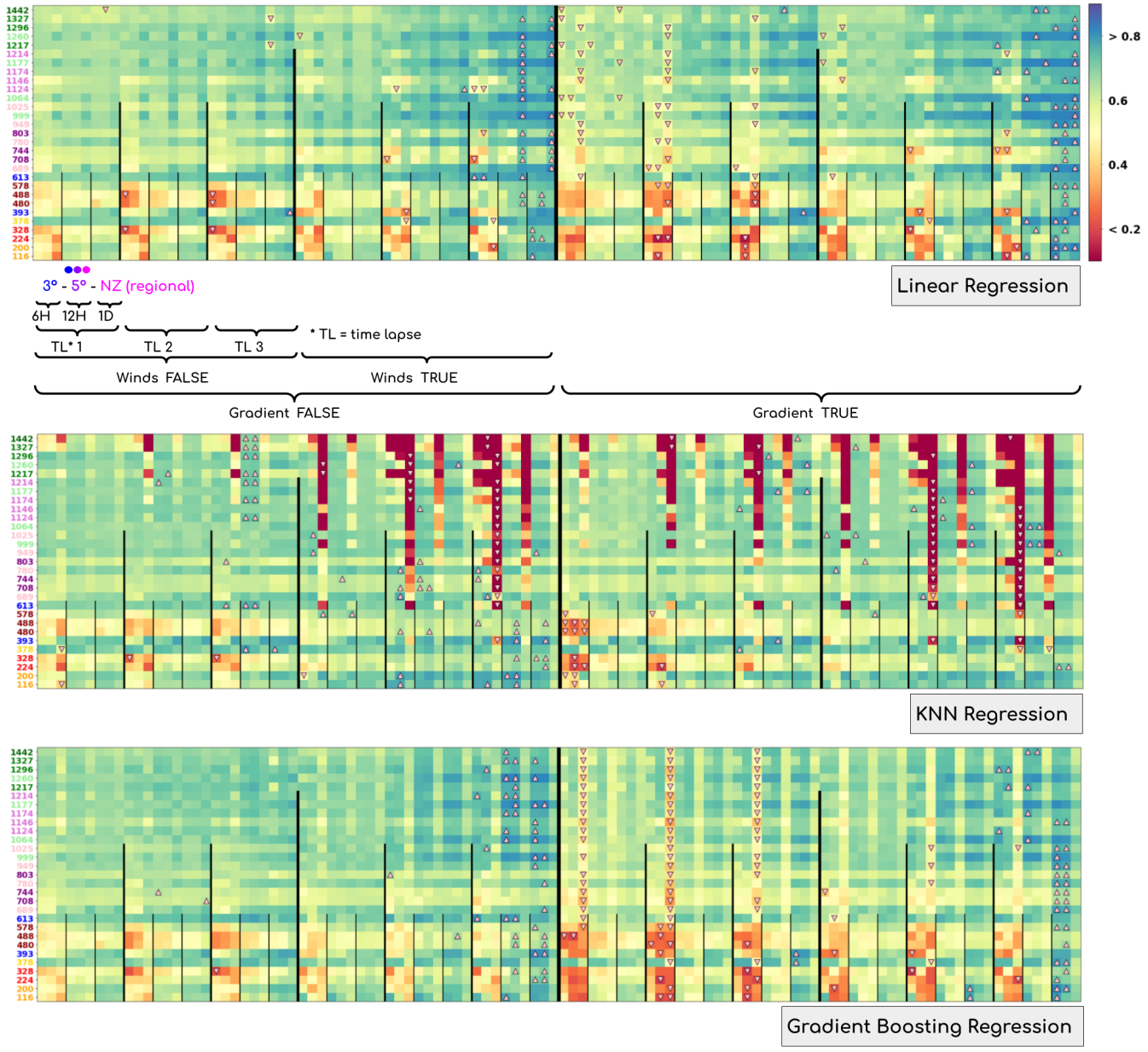

Figure 5. Models performance for all the available predictors. The distribution of the features for each predictor is described in the bottom of the top subplot, and the black lines help visualize the frontiers between the different predictors too. The metric plotted is the KGE' in all the cases, and as it can be seen in the different subplots, there are predictors which always behave better than others, as there are also locations where the behaviour of the storm surge is also more predictable. For the up and down triangles appearing in each row, they represent the best and worst 3 performing predictors, respectively.

4.1.1 Experiments general explanation

On the one hand, linear regression always behaves better as more data are provided as input, achieving the best results when daily and regional predictors are used, and the prediction worsens as we decrease the amount of information fed into the model. On the other hand, k-NN and gradient boosting regressions, even though it is clear that they also require sufficient information to behave well, perform worse in situations where the regional predictor is used, but work better when local features influencing the storm surge are utilized (outperforming linear regression in these cases, see Fig.6). Examples of these features are the projected winds and the spatially local predictors (see how well, for example, the gradient boosting regressor performs with local predictors, using the winds in 12 and 24 hourly resampled data; last columns in the first half of the XGB-subplot in the bottom of Fig.5). In relation to the temporal memory of the predictor, experiments show it is always best to have, at least, the information of the 2 previous time steps to the moment from the prediction/reconstruction.

It is interesting how not including the gradients seems to have a positive impact on k-NN and gradient boosting but the reverse is true for the linear regression case, specially when using the regional predictor. Regarding the projected winds behaviour, the 3 plots in Fig.5 suggest that they outperform the function of the gradients, reaching better KGE' values for almost all the scenarios, and for the 3 statistical models (here we compare the region in the middle of the plots, so Gradients=False and Winds=True, with Gradients=True and Winds=False).

Moreover, the time resample parameter is also crucial in predictor and model selection, as it clearly affects storm surge reconstructions. With the 6 hourly resampled data, predictors behave very poorly, but in the case of the k-NN, when selecting a very low number of neighbors (see Fig.7), there exist cases when this predictor gets good model performances (see the low part at the end of the first half of the k-NN plot in Fig.5). For the other 2 scenarios, so 12 and 24 hourly resampled data, models usually perform relatively well, reaching values above 0.6 in most of the cases.

Finally, a clear pattern can be seen between the two islands of NZ (better seen in Fig.8), nodes located to the north are more predictable than southern regions, and the west coast always behaves better than the east one. This is explained by the fact that those coasts are exposed to the dominant storm direction, which sees low pressure systems traveling from south Australia to reach the south - south-western coast of New Zealand (Stephens et al., 2019b) (notice here that tropical-cyclon events are not very well represented in the data, and future works will include these storms, see Hodges et al. (2017)). This reasoning might also explain why the locations in the east of the south island exhibit bad results when working with 6H predictors, as they might be more sensitive to past conditions, and the 6H predictors contain very little of this information.

Once the relationship between the predictor, the models and the location around NZ is understood, we now focus on the influence of the hyperparameters of the k-NN and gradient boosting regressors on model performance. For the gradient boosting model, we tried different tree maximum depths and minimum number of data points per final leaf on the decision tree, and experiments show an optimal convergence at depth$\approx$18 and when at least 4% of the data at each final leave is considered. When the tree is more pruned and leaves start acquiring fewer data than this, the model might start overfitting.

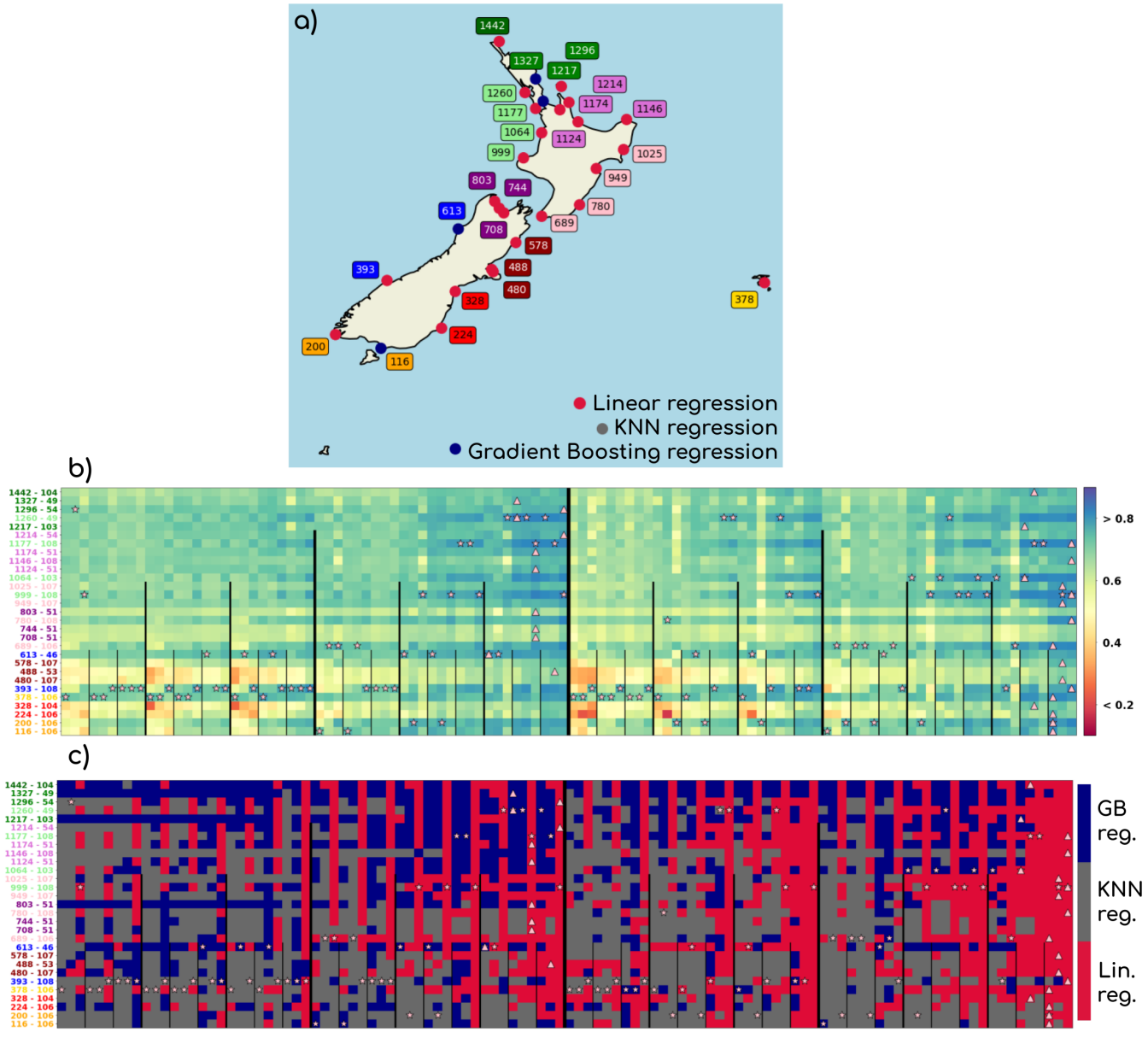

Figure 6. a) Shows the best performing statistical model for each studied location. In b), the KGE' values are shown and in c), the best statistical model. Moreover, b) and c) shows the winning predictor for each location (up arrows), and the winning location for each predictor (stars).

In the case of k-nearest neighbors, the most informative parameter is the number of neighbors used, which can also help understanding the results presented above (please refer to Figure 7). The selection of this hyperparameter is not straightforward and can lead to overfitting problems too. This is known in machine learning as the bias-variance trade-off.

On the one hand, when the number of neighbors is very low, or even 1, the bias in the reconstructed data is minimal, as the closest data point in a dataset of around 10,000 (if data is resampled to daily maxima) rows, with a relatively low standard deviation, is usually very similar to the real data point studied (more if the predictor is local). But the overall variance in the data is large, besides, if the historical time used to calibrate the data differs from the testing time, then overfitting problems will appear undoubtedly. On the other hand, if the model is trained with a bigger number of neighbors, the bias in the results will become also higher, but the variance will be reduced, and so the capacity to predict extreme events, which in this case can be flooding cases.

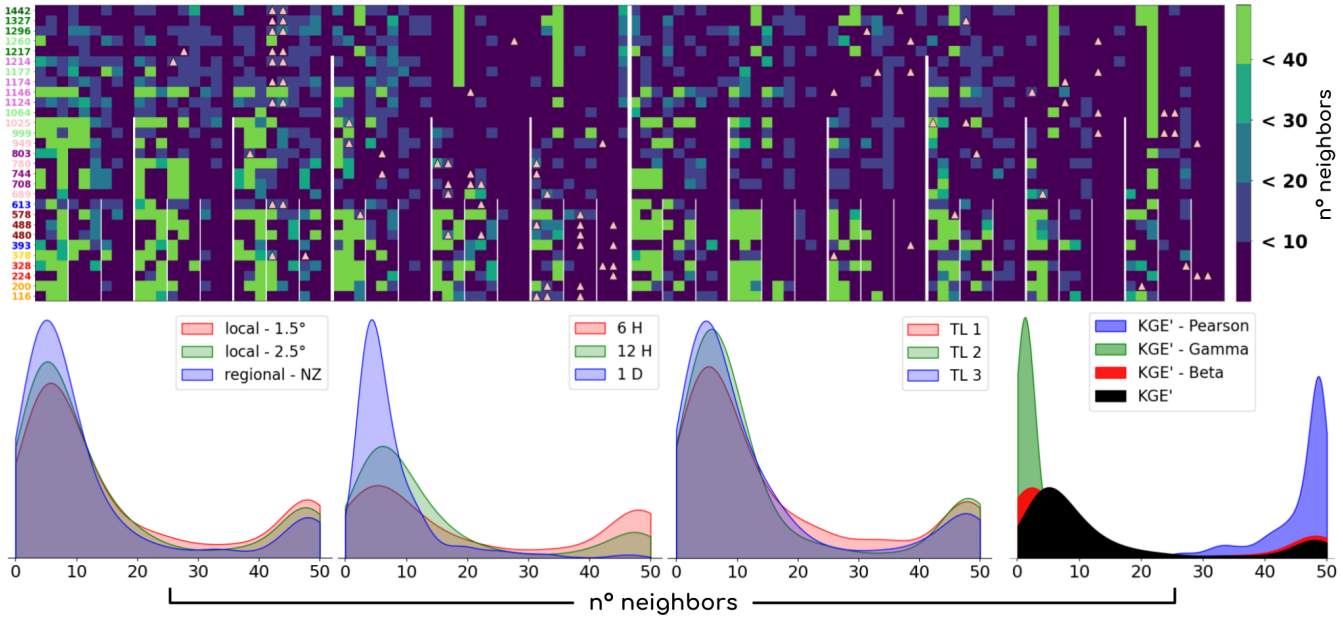

Figure 7. The best number of neighbors found is shown for each combination of predictor and location. The density plots below represent how different predictors might influence the number of neighbors that is appropriate for the model to work. Finally, in the right-down side, the KGE' is broken down in its three components.

As in this study we evaluate the model performances using KGE', which has the capacity to evaluate how the model behaves in the extremal regions of the data (Kling et al., 2012); we can see in Fig.7 how well performing models present a lower number of neighbors, while predictors with low amounts of information, which is the case of the 6 hourly ones, present higher number of this hyperparameter, due to the lack of information in the very closest neighbors. As a summary, we could say the inherent bias-variance trade-off of the k-NN model deals this time with the amount of information each predictor has, leading to complex but highly informative results.

In addition, the subplot located in the low-right part of Fig.7 shows how the 3 components of the KGE' metric behave, finishing with the final density plot of the total metric in black. The γ coefficient, which inspects the relationship between the standard deviation and the mean in the reconstructed and real values in the dataset, has bigger values when very low number of neighbors is chosen, as the β coefficient does, which can be interpreted as the bias term (remember the bias-variance trade-off), even though the density plot is not that abrupt here. However, the pearson correlation coefficient always performs better close to the highest number of neighbors, as it measures the temporal dynamics of the whole population, which is very influenced by the mean values.

4.1.2 Best performing predictor selection

Determining which is the best predictor overall is not a straightforward task. As we have just demonstrated, predictor performance is highly influenced by the model and the data that is fed into it, so depending on the purpose of the reconstruction, specially its time resolution, this optimal predictor might change.

The atmospheric predictor that has been chosen as the best to carry out the reconstruction of the historical storm surge maxima along the New Zealand coastline can be found at the right of the figures (the penultimate column), where all the variables are used; the gradient, the projected winds, time lapse of 3 (the day and two days before prediction) and daily resampling, but the local predictor of 5 by 5 degrees, even though the regional one seems to be working better in some of the cases.

The reasons we are using this local predictor are related to the fact that the k-NN and the gradient boosting regressions behave better with local features. Storm surge is a long wave, and it depends on the atmospheric situation over a large spatial domain (Bell et al., 2000), but providing the model too much information might decrease its performance (see Hastie et al. (2001) and Friedman (2000) for more insights in how a large amount of features might decrease model performance). Besides, computational efforts decrease with local predictors. We included the daily temporal resolution predictor in our analysis to allow comparison with other studies (it is the most common time resolution used, and models behave better in this case too). However, our study might change this approach.

After this summary of the results extracted from the experiments analysis, we will now focus in comparing the spatial pattern around the New Zealand coastline, visualizing how the best predictor performs for the different models tested.

4.2 New Zealand coastline reconstruction

Now the best predictor has been identified, for which all the variables are used; the gradient, the projected winds, time lapse of 3 and daily resampling, but the local predictor of 5 by 5 degrees, this predictor is used in all the models to reconstruct the historical storm surge daily maximum values in the entire coast of New Zealand. Results for the KGE' metric are shown in Figure 8.

In agreement with the results presented in the earlier sections, the regions which are more exposed to storms coming from the south of Tasmania are more predictable in terms of the storm surge maxima and present better results, while regions which are hidden from this dominant storms might be more difficult to predict.

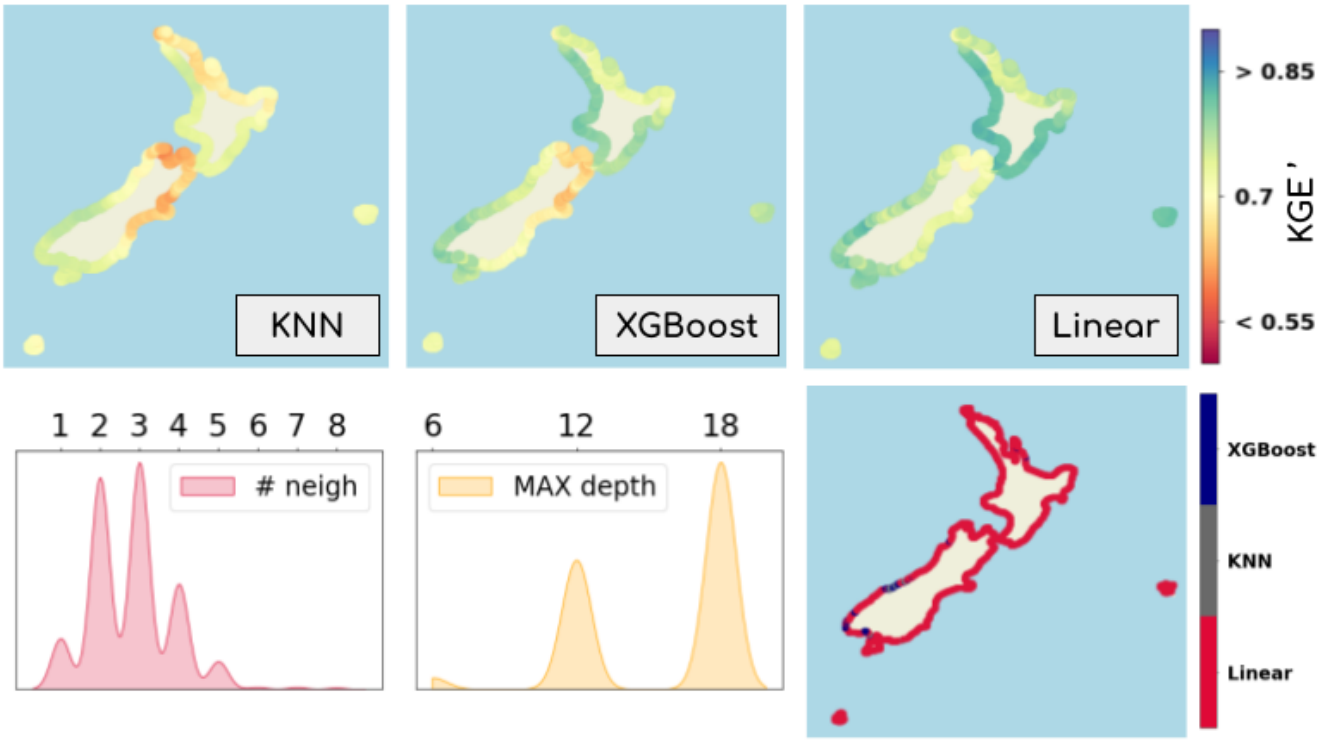

Figure 8. The Modified Kling-Gupta Efficiency is shown for all the nodes of the Moana v2 hindcast, for all the statistical models used, and for the best found predictor. The kde plots provide information about which hyperparameters achieve the better results for each model. Finally, the best statistical model is also shown for each hindcast node.

Linear regression is the best performing model in the majority of the cases, been just improved by gradient boosting in a few cases in well predicted regions. k-nearest neighbors is not able to improve these two models performances in any case for this best behaving predictor.

This better performance of the linear regression technique was explained before in the experiments results, and we can conclude that highly informative atmospheric predictors work better with multi-linear regression. However, we can see how k-NN and gradient boosting underperform, although gradient boosting gets very close results to its principal competitor, the multi-linear regression. In the case of k-NN, where the best results are obtained with a very low number of neighbors, results are always below the other statistical methods.

The 2 kde plots in the middle of Fig.8 explain how the hyperparameters in the models are chosen. For the k-NN case, a very low number of neighbors is usually chosen as best by the KGE' metric, as enough data is available, and for the gradient boosting regression, the kde has it high in 18 maximum depth, which represent the more complex tree.

5 Discussion

In this article we have studied the relationships between the possible combinations of atmospheric predictor, statistical model and the area of New Zealand in which they have been applied. If we call each combination an experiment, we evaluate a total of more than 9000 experiments (many more if we count the repeated experiments depending on the different hyperparameters used), in order to reach the conclusions outlined below.

In addition, the behaviour of all the statistical models was evaluated and the winning predictor is used to determine the predictability of the maximum value of storm surge, resolved by location, in NZ. Thus, given this best predictor, multi-linear regression is the model that achieves the highest values for our metric of choice, which is the Modified Kling-Gupta Efficiency statistic. It is proposed in Kling et al. (2012) and is capable of accurately measuring different components of interest in any temporal regression problem, allowing particular focus on the behaviour of the reconstruction in the extremes.

The following are the main conclusions drawn from this study:

- **(predictors)** The main feature observed is in relation to the atmospheric predictor, and is the trade-off between the temporal resolution (resampling of the data) and the amount of information that the proposed methodology is capable of processing. Thus, the hourly predictors, although providing more recent information for the time when the prediction is made, only manage to provide, at most, information for the last 18 hours. However, it has been demonstrated here that the storm surge can be influenced by the atmospheric conditions existing one or even two days before the time of analysis. In this way, the predictors that have generally achieved better results are the daily predictors, incorporating the two natural days prior to the reconstruction, and when both the projected winds and the sea-level-pressure gradients are used in the building of the predictor.

- **(statistical models)** Secondly, the behaviour of each statistical model is different, depending on the amount and variety of information provided on atmospheric conditions. The multi-linear regression model, which has outperformed the other two linear models, k-NN and gradient boosting, for most daily and regional predictors, increases its accuracy whenever more info is added to the predictor, but this does not happen with the other two models. In cases where more model expertise is required, either by increasing the temporal resolution of the reconstruction, or by a spatial reduction of the predictor, the k-NN and gradient boosting models improve the linear regression expertise by a fairly high percentage (see Fig.5), which is a key conclusion of the study. The reason for this behaviour might be the large number of principal components that is retained with the daily and regional predictors, see Wilks (2005) for a detailed explanation on how a bigger number of independent variables might affect the dependent variable reconstruction. In the k-NN model, this high number of PCs means that the model has to go to a very large number of neighbors to work without overfitting, not performing well, and in the case of trees (gradient boosting), the selection process to bifurcate the tree becomes very costly and imprecise, as the feature engineering step is very important here (Hastie et al., 2001; Friedman, 2000). This is not the case of smaller (local of 3·3 or 5·5 degrees) predictors, having these the local atmospheric behaviours more represented in the first PCs.

– **(location)** The spatial pattern of performance observed is clear for all models; the areas more exposed to storms coming from the south-west (which are the more predominant ones, see Stephens et al. (2019b)) present a more predictable storm surge, while those areas more sheltered from the same storms are harder to predict, as the storm has already interacted with the land by the time it reaches these locations and the surge has interacted with the coastline so that it is a much more complex situation. This task is even more challenging in areas with high variations of storm surge, as is generally the southern island, and using predictors that in themselves contain less total information, such as, for example, the 6-hourly predictors.

Now that an understanding of the atmospheric conditions affecting the storm surge behaviour has been developed, as well as the performance of the different statistical models, the insight gained in this study can be used to train new statistical models that are able to exploit the non-linear relationships between the predictor and the predictand. In this sense, hybrid models might offer very interesting solutions, while the promising neural networks, offering innovative architectures, might also contribute to improve these results.

6 Conclusions

A methodology has been proposed in this article that is capable of reconstructing the storm surge maximum levels over the entire coast of New Zealand. First, we have validated the existing sea level hindcast with observational tidal gauges and then, the best atmospheric predictor has been found. In this study, we have used 3 different linear models to perform our analysis, which are multi-linear regression, k-NN regression and gradient boosting regression. In this way, we have demonstrated that the more data the atmospheric predictor has, the better these models work, but spatially talking, the increment of the spatial resolution in the data might wrongly affect the performance of some of the models. Finally, and given a local predictor, the storm surge time series in New Zealand have been daily reconstructed, improving previous studies and obtaining excellent results for the KGE' statistic and some other regression metrics that can be found in the GitHub repository of the project.

Future work will involve testing different encoding techniques, more complex than Principal Components Analysis, that might help alleviate some of the limitation imposed by the hardware requirements of PCA as the number of features increases. Fortunately, showing that using a local predictor lead to the best overall performance, means that this might not be as much of an issue. Moreover, non-linear models will be trialled to reconstruct the storm surge maximum levels. To this end, neural networks might seem appropriate to both encode the data and perform the storm surge reconstruction, as suggested in Tiggeloven et al. (2021) and Bruneau et al. (2020), but hybrid models can also be a promising solution.

Other possible lines of research involve changing the way in which the problem has been posed, since right now, if one wants to predict the value of the storm surge for the next few days, it is necessary to have values of the atmospheric conditions for the next few days. Future models could be trained to predict future values of the storm surge, given the atmospheric history, without the need to use atmospheric forecasts, but rather the real value of the atmosphere at that moment.

Acknowledgements. The authors would like to acknowledge the National Centers for Environmental Prediction for generating the atmospheric data used in this study, the MetService in New Zealand for producing the storm surge hindcast and the New Zealand councils for providing the tidal gauges data. This study is funded by the New Zealand Ministry of Business Innovation and Employment, under contract number MSVC1901.

References

Altman, N. S.: An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression, The American Statistician, 46, 175–185, http://www.jstor.org/stable/2685209, 1992.

Bell, R., Goring, D., and de Lange, W.: Sea-level change and storm surges in the context of climate change, Institution of Professional Engineers New Zealand Transactions, 27, 1–10, 2000.

Bell, R. G. and Goring, D. G.: Techniques for analyzing sea level records around New Zealand, Marine Geodesy, 19, 77–98, https://doi.org/10.1080/01490419609388071, 1996.

Brenstrum, E.: The cyclone of 1936: the most destructive storm of the Twentieth Century?, Weather and Climate, 20, 23–27, 2000.

Bruneau, N., Polton, J., Williams, J., and Holt, J.: Estimation of global coastal sea level extremes using neural networks, Environmental Research Letters, 15, 074 030, https://doi.org/10.1088/1748-9326/ab89d6, 2020.

Cagigal, L., Rueda, A., Castanedo, S., Cid, A., Perez, J., Stephens, S. A., Coco, G., and Méndez, F. J.: Historical and future storm surge around New Zealand: From the 19th century to the end of the 21st century, International Journal of Climatology, 40, 1512–1525, https://doi.org/https://doi.org/10.1002/joc.6283, 2020.

Camus, P., Méndez, F. J., Losada, I. J., Menéndez, M., Espejo, A., Pérez, J., Rueda, A., and Guanche, Y.: A method for finding the optimal predictor indices for local wave climate conditions, Ocean Dynamics, 64, 1025–1038, https://doi.org/https://doi.org/10.1007/s10236-014-0737-2, 2014.

Cid, A., Camus, P., Castanedo, S., Méndez, F. J., and Medina, R.: Global reconstructed daily surge levels from the 20th Century Reanalysis (1871–2010), Global and Planetary Change, 148, 9–21, https://doi.org/https://doi.org/10.1016/j.gloplacha.2016.11.006, 2017.

Cid, A., Wahl, T., Chambers, D. P., and Muis, S.: Storm Surge Reconstruction and Return Water Level Estimation in Southeast Asia for the 20th Century, Journal of Geophysical Research: Oceans, 123, 437–451, https://doi.org/https://doi.org/10.1002/2017JC013143, 2018.

Codiga, D.: Unified tidal analysis and prediction using the UTide Matlab functions, https://doi.org/10.13140/RG.2.1.3761.2008, 2011.

de Lange, W. and Gibb, J.: Seasonal, interannual, and decadal variability of storm surges at Tauranga, New Zealand, New Zealand Journal of Marine and Freshwater Research - N Z J MAR FRESHWATER RES, 34, 419–434, https://doi.org/10.1080/00288330.2000.9516945, 2000.

de Souza, J. M. A. C.: Moana Ocean Hindcast, https://doi.org/10.5281/zenodo.5895265, 2022.

Fix, E. and Hodges, J. L.: Discriminatory Analysis - Nonparametric Discrimination: Consistency Properties, International Statistical Review, 57, 238, 1989.

Friedman, J.: Greedy Function Approximation: A Gradient Boosting Machine, The Annals of Statistics, 29, https://doi.org/10.1214/aos/1013203451, 2000.

Goring, D. G. and Bell, R. G.: Distilling information from patchy tide gauge records: The New Zealand experience, Marine Geodesy, 19, 63–76, https://doi.org/10.1080/01490419609388070, 1996.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, Journal of Hydrology, 377, 80–91, https://doi.org/https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Gutiérrez, J., Ancell, R., Cofiño, A., and Sordo, C.: Redes Neuronales y Probabilísticas en las Ciencias Atmosféricas, 2004.

Haidvogel, D., Arango, H., Budgell, W., Cornuelle, B., Curchitser, E., Di Lorenzo, E., Fennel, K., Geyer, W., Hermann, A., Lanerolle, L., Levin, J., McWilliams, J., Miller, A., Moore, A., Powell, T., Shchepetkin, A., Sherwood, C., Signell, R., Warner, J., and Wilkin, J.: Ocean

forecasting in terrain-following coordinates: Formulation and skill assessment of the Regional Ocean Modeling System, Journal of Computational Physics, 227, 3595–3624, https://doi.org/https://doi.org/10.1016/j.jcp.2007.06.016, predicting weather, climate and extreme events, 2008.

Hastie, T., Tibshirani, R., and Friedman, J.: The Elements of Statistical Learning, Springer Series in Statistics, Springer New York Inc., New York, NY, USA, 2001.

Heath, R. A.: Significance of storm surges on the New Zealand coast, New Zealand Journal of Geology and Geophysics, 22, 259–266, https://doi.org/10.1080/00288306.1979.10424224, 1979.

Hodges, K., Cobb, A., and Vidale, P.: How Well Are Tropical Cyclones Represented in Reanalysis Datasets?, Journal of Climate, 30, 5243–5264, https://doi.org/10.1175/JCLI-D-16-0557.1, 2017.

Kling, H., Fuchs, M., and Paulin, M.: Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios, Journal of Hydrology, 424-425, 264–277, https://doi.org/https://doi.org/10.1016/j.jhydrol.2012.01.011, 2012.

Nash, J. and Sutcliffe, J.: River flow forecasting through conceptual models part I — A discussion of principles, Journal of Hydrology, 10, 282–290, https://doi.org/https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

Rueda, A., Cagigal, L., Antolínez, J. A. A., Albuquerque, J. C., Castanedo, S., Coco, G., and Méndez, F. J.: Marine climate variability based on weather patterns for a complicated island setting: The New Zealand case, International Journal of Climatology, 39, 1777–1786, https://doi.org/https://doi.org/10.1002/joc.5912, 2019.

Saha, S., Moorthi, S., Pan, H.-L., Wu, X., Wang, J., Nadiga, S., Tripp, P., Kistler, R., Woollen, J., Behringer, D., Liu, H., Stokes, D., Grumbine, R., Gayno, G., Wang, J., Hou, Y.-T., Chuang, H.-Y., Juang, H.-M. H., Sela, J., Iredell, M., Treadon, R., Kleist, D., Delst, P. V., Keyser, D., Derber, J., Ek, M., Meng, J., Wei, H., Yang, R., Lord, S., van den Dool, H., Kumar, A., Wang, W., Long, C., Chelliah, M., Xue, Y., Huang, B., Schemm, J.-K., Ebisuzaki, W., Lin, R., Xie, P., Chen, M., Zhou, S., Higgins, W., Zou, C.-Z., Liu, Q., Chen, Y., Han, Y., Cucurull, L., Reynolds, R. W., Rutledge, G., and Goldberg, M.: NCEP Climate Forecast System Reanalysis (CFSR) Selected Hourly Time-Series Products, January 1979 to December 2010, https://doi.org/10.5065/D6513W89, 2010.

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., Behringer, D., Hou, Y.-T., ya Chuang, H., Iredell, M., Ek, M., Meng, J., Yang, R., Mendez, M. P., van den Dool, H., Zhang, Q., Wang, W., Chen, M., and Becker, E.: NCEP Climate Forecast System Version 2 (CFSv2) Selected Hourly Time-Series Products, https://doi.org/10.5065/D6N877VB, 2011.

Siek, M.: Predicting Storm Surges: Chaos, Computational Intelligence, Data Assimilation and Ensembles: UNESCO-IHE PhD Thesis, CRC Press, 2019.

Stephens, S., Coco, G., and Bryan, K.: Numerical Simulations of Wave Setup over Barred Beach Profiles: Implications for Predictability, Journal of Waterway Port Coastal and Ocean Engineering, 137, https://doi.org/10.1061/(ASCE)WW.1943-5460.0000076, 2011.

Stephens, S., Bell, R., and Haigh, I.: Spatial and temporal analysis of extreme sea level and skew surge events around the coastline of New Zealand, https://doi.org/10.5194/nhess-2019-353, 2019a.

Stephens, S., Bell, R., and Haigh, I.: Spatial and temporal analysis of extreme sea level and skew surge events around the coastline of New Zealand, https://doi.org/10.5194/nhess-2019-353, 2019b.

Stephens, S., Bell, R., and Haigh, I.: Spatial and temporal analysis of extreme storm-tide and skew-surge events around the coastline of New Zealand, Natural Hazards and Earth System Sciences, 20, 783–796, https://doi.org/10.5194/nhess-20-783-2020, 2020.

Tadesse, M. G. and Wahl, T.: A database of global storm surge reconstructions, Scientific Data, 8, 125, https://doi.org/10.1038/s41597-021-00906-x, 2021.

Thomson, R. E. and Emery, W. J.: Data Analysis Methods in Physical Oceanography, Elsevier, Boston, third edition edn., https://doi.org/https://doi.org/10.1016/B978-0-12-387782-6.05001-8, 2014.

Tiggeloven, T., Couasnon, A., van Straaten, C., Muis, S., and Ward, P. J.: Exploring deep learning capabilities for surge predictions in coastal areas, Scientific Reports, 11, 17 224, https://doi.org/10.1038/s41598-021-96674-0, 2021.

Wang, X., Swail, V., and Cox, A.: Dynamical versus statistical downscaling methods for ocean wave heights, International Journal of Climatology, 30, 317 – 332, https://doi.org/10.1002/joc.1899, 2009.

Wikipedia: Gauss-Markov theorem, https://en.wikipedia.org/wiki/Gauss%E2%80%93Markov_theorem.

Wilks, D.: Statistical Methods in the Atmospheric Sciences, Volume 91, Second Edition (International Geophysics), 2005.

Appendix A: Principal Components Analysis (PCA)

The predictor matrix has as many rows as times or entries in the dataset, about 10,000 in the case of a resampling to daily maxima, and as many columns as variables have been added. Notice here that each coordinate of each atmospheric variable is a scalar feature to be analysed with the Principal Components Analysis (Gutiérrez et al., 2004; Hastie et al., 2001).

Then the analysis is performed, and the atmospheric predictor is projected in a new space, where the first variables in this new space explain the highest percentage of the variance in the data. This data transformation is just an orthogonal linear transformation that converts the data to a new coordinate system such that the greatest variance by some scalar projection of the data comes to lie on the first coordinate (called the first principal component, PC_1), the second greatest variance on the second coordinate, PC_2, and so on... where the new weights that will transform the original atmospheric predictor into the new basis are represented in the equation below:

$$\mathbf{w} = \frac{\mathbf{w}\mathbf{X}^T\mathbf{X}\mathbf{w}}{\mathbf{w}^T\mathbf{w}} \rightarrow w_{(1)} = \arg\max\left\{\frac{\mathbf{w}\mathbf{X}^T\mathbf{X}\mathbf{w}}{\mathbf{w}^T\mathbf{w}}\right\} \tag{A1}$$

where $\mathbf{X}^T\mathbf{X}$ is the covariance matrix of the original atmospheric data (bold variables imply they are matrices). Once the new variables are obtained, only the ones representing a desired percentage of the variance are selected to be used by the statistical models. Here we choose 98% of the variance, allowing to reduce the dimensionality of the data, which is very large when working with ocean and atmosphere datasets, while keeping a large proportion of the total information.

When the new variables (PCs) and their physical weights (EOFs) are obtained, the original data can be recalculated using the following expression below:

$$\mathbf{X}(x,t_i) = \mathbf{EOF}_1(x) \times \mathrm{PC}_1(t_i) + \mathbf{EOF}_2(x) \times \mathrm{PC}_2(t_i) + ... + \mathbf{EOF}_n(x) \times \mathrm{PC}_n(t_i) \tag{A2}$$

where the PCs represent the contribution of each EOF in time, and the EOFs represent the oscillation modes that together reconstruct the physically understandable variables (see Fig.1). With this PC analysis, the data is reprojected in a easily reducible new space, where the first variables in this new space have two principal characteristics: they represent the higher percentage of the variance in the original data (1) and they have physical meaning (2), see Camus et al. (2014). For a more mathematical description, please see Hastie et al. (2001), and check Gutiérrez et al. (2004) to discover PCA case studies in atmospheric sciences.

Appendix B: Statistical models detailed explanation

B1 Linear regression

In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables, respectively). The case of one explanatory variable

is called simple linear regression; for more than one, the process is called multiple linear regression (multi-linear). This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data (this is a machine learning strategy, Hastie et al. (2001)). Such models are called linear models. Most commonly, the conditional mean of the response (or predictand) given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used (see generalized linear models, GLMs at Hastie et al. (2001)). Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis.

For our case study, the dependent variable (storm surge maxima) will be reconstructed given a set of independent variables (atmospheric predictor), and the β coefficients in the equation below represent the coefficients that are inferred from the data:

$$y = \mathbf{X}\beta + \varepsilon \text{ , where: } y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} \mathrm{x}_1^T \\ \mathrm{x}_2^T \\ \vdots \\ \mathrm{x}_n^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}. \tag{B1}$$

where p relates to the number of independent variables (columns in the 2-dimensional dataset) and n is the number of data points or rows in the dataset. Given this equation, the optimal coefficients can be calculated so the quadratic sum of errors is minimized:

$$\beta_{opt} = \arg\min_{\beta} \sum_{i=1}^{n} (\beta \cdot x_i - y_i)^2 \rightarrow \beta_{opt} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y \tag{B2}$$

where once again, the index i refers to each data point or row in the data. Notice that to prove that the β obtained is indeed the local minimum, one needs to differentiate once more to obtain the Hessian matrix and show that it is positive definite. This is provided by the Gauss–Markov theorem, see Hastie et al. (2001) and Wikipedia.

The principal benefit of this methodology is clear given its definition, as long as its coefficients are linear, the optimal coefficients are deterministic, leaving no room to convergence issues. Moreover, the time this model takes to calculate these coefficients is minimal, although it requires a sufficient RAM capacity to invert the big data matrices.

B2 *k*-NN regression

Another machine learning tool used is the k-nearest-neighbors algorithm (k-NN), which is a non-parametric classification/regression method first developed by Evelyn Fix and Joseph Hodges in 1951, see Fix and Hodges (1989), and later expanded by Thomas Cover, see Altman (1992).

In k-NN regression the output is the property value for the object. This value is the average of the values of k-nearest neighbors (although also a weighted sum might be applied). k-NN is a type of regression where the function is only approximated locally and all computation is deferred until function evaluation. Since this algorithm relies on distance, if the features represent different physical units or come in vastly different scales then normalizing the training data can improve its accuracy dramatically, but in this case, the principal components are not normalized, as usually the first PCs have values much larger than the following ones, and are just these ones which we are more interested in.

In this line, the training examples are vectors in a multidimensional feature space, each with a predicted scalar value. The training phase of the algorithm consists only in storing the feature vectors and predicted values of the training samples. Then, the validation process will evaluate how the reconstruction of the test data differs from the original storm surge values, where the k-NN will be the nearest neighbors to the new data point, with respect to a distance metric, which in our case is the euclidean distance. This model has different hyperparameters that can be fine-tuned, as the already mentioned distance metric, which might indeed require several parameters, or the weights each closest neighbor might get to calculate the final weighted value, if this is required, but the hyperparameter with the most influence in the model performance is the number of neighbors used (Altman, 1992; Fix and Hodges, 1989; Gutiérrez et al., 2004).

The number of neighbors is crucial, and depending on the problem, the way to find the optimal value of this parameter might differ. In this study, we evaluate the models performance with the Modified Kling-Gupta Efficiency, see Kling et al. (2012), explained in the results section, as the behavior of our models in the extreme values is very important. In this context, the less neighbors are chosen, the better the model reconstructs the extreme events in the historical dataset, but more bias is introduced in the predictions. This is known as the bias-variance trade-off, and is one of the main concepts of discussion in machine learning nowadays.

B3 Gradient boosting regression

Another used statistical method is gradient boosting, which builds an additive model in a forward stage-wise fashion that allows for the optimization of an arbitrary differentiable loss function (usually the mean squared error, which is usually divided by two to simplify its derivative), see Friedman (2000). At each stage, a regression tree is fitted on the negative gradient of the given loss function, so it exploits the capabilities of several regression trees together, but using them so a loss function is optimized, and then the final output is optimal.

Gradient boosting is a machine learning technique used in regression and classification tasks. It gives a prediction model in the form of an ensemble of weak prediction models (ensemble methods), which are typically decision trees. When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms random forest and Adaboost, see Friedman (2000), where the ensemble is only used to calculate the final mean off all the pre-trained trees, and the weak learners are added in series, respectively.

To understand gradient boosting, decision trees must be also explained. Tree-based methods partition the feature space into a set of rectangles, and then fit a simple model (like a constant) in each one. They are conceptually simple yet powerful. Let's consider a regression problem with continuous response y, which is the storm surge signal, and inputs $\mathbf{X}$, which are the PCs of

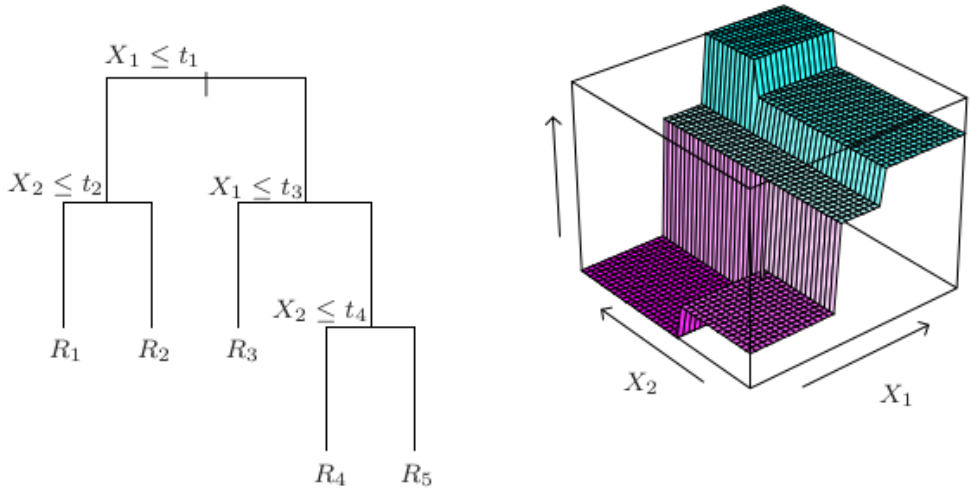

Figure B1. The final structure of an individual "weak" learner / decision tree is shown. In the left, the partitions depending on two input variables are shown, and in the right, the final $f(\mathbf{X})$ surface, reconstructing the target values depending on X_1 and X_2 can be also depicted.

the atmospheric predictor. Figure B1 shows a partition of the feature space where just the first and the second PCs are taken into account, and how the final $f(\mathbf{X})$ should look like if the mean value at each final leaf is calculated. This partitions of the feature space are calculated so this previously mentioned metric is minimized, although different criterions might be used.

As it is shown in the figure, the output of the model can be explained with Eq.B3:

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m) \tag{B3}$$

where the output of the model for each point x will be the mean of all the target values of the group x belongs to, if the mean squared error divided by 2 is minimized, which is the case.

Now, like other boosting methods, gradient boosting combines these weak "learners" into a single strong learner in an iterative way. It is easiest to explain in the least-squares regression setting, where the goal is to "teach" a model to predict values of the form $\hat{y} = f(x)$ by minimizing the mean squared error $\frac{1}{p}\sum_i (\hat{y}_i - y_i)^2$, where i indexes over some training set of size n of actual values of the output variable y, i.e., the storm surge.

Now, let us consider a gradient boosting algorithm with M stages / iterations. At each stage m of gradient boosting, suppose some imperfect model F_m (for low m, this model may simply return $\hat{y}_i = \bar{y}$). In order to improve F_m, our algorithm should add some new estimator, $h_m(x)$. Thus,

$$F_{m+1}(x) = F_m(x) + h_m(x) = y \text{ or } h_m(x) = y - F_m(x) \tag{B4}$$

Therefore, gradient boosting will fit h to the residuals. As in other boosting variants, each F_{m+1} attempts to correct the errors of its predecessor F_m. Finally, we end up with M decision trees, which as a group, outperforms the capabilities of individual "weak" trees.