An integrated framework that combines machine learning and numerical models to improve wave-condition forecasts

Fearghal O'Donncha[a,*], Yushan Zhang[b], Bei Chen[a], Scott C. James[c]

[a]*IBM Research - Ireland, Dublin*
[b]*University of Notre Dame*
[c]*Baylor University*

Abstract

This study investigates near-shore circulation and wave characteristics applied to a case-study site in Monterey Bay, California. We integrate physics-based models to resolve wave conditions together with a machine-learning algorithm that combines forecasts from multiple, independent models into a single "best-estimate" prediction of the true state. The Simulating WAves Nearshore (SWAN) physics-based model is used to compute wind-augmented waves. Ensembles are developed based on multiple simulations perturbing data input to the model. A learning-aggregation technique uses historical observations and model forecasts to calculate a weight for each ensemble member. We compare the weighted ensemble predictions with measured data to evaluate performance against present state-of-the-art. Finally, we discuss how this framework that integrates data-driven and physics-based approaches can outperform either technique in isolation.

1. Introduction

Physics-based numerical models are defined by: (1) the physical formulation, (2) numerical discretization, and (3) input data driving the simulations. Typically, all three involve some degree of uncertainty. Wave modeling

*Corresponding author

Email address: `feardonn@ie.ibm.com` (Fearghal O'Donncha)

and predictions result from solution of the spectral-action balance equations, which are based on an approximation of reality derived from an incomplete data set (Komen et al., 1996; Mei et al., 1989). Similarly, numerical discretization approximates the solution to these equations with accuracy dependent upon spatial resolution and time steps. In addition, wave-condition forecasting involves multicomponent input data such as bathymetry, ocean characteristics, and meteorological information. Rogers et al. (2005) observed that in global operational nowcast/forecast systems, wind forcing is a dominant source of error (in global wave models the only time-varying inputs are typically wind forcing). Here, we combine ensemble-forecasting and machine-learning techniques to: (1) investigate uncertainty from an advanced wave-modeling package and (2) generate a forecast that is better than the best individual model prediction.

Ensemble-based machine-learning approaches (Mallet et al., 2009) comprise aggregate ensemble predictions based on multiple simulations where anything from physical parameterizations, numerical discretization, or input data are perturbed. The learning-aggregation technique presented here, makes use of historical observations and model forecasts to produce a weight for each model. A linear or convex (i.e., where weights sum to 1) combination of model forecasts is performed with these weights to generate a best model forecast. A key component in wave forecasting is representation of extreme events, an area where traditional machine learning under-performs because such algorithms typically depend upon predicting the conditional mean of the data. By combining physical models with machine learning, we propose to overcome this shortcoming.

Ensembles forecasts of wave conditions are typically generated from statistical perturbations of wave-height boundary data, ocean-current input data, wind forcing (particularly for global models), model physics, discretization, or parameterization schemes (Chen, 2006). The fundamental objective of ensemble forecasting or prediction is to investigate the uncertainty inherent in forecasting to provide more information about future states. This process facilitates transition from single, deterministic forecasting with optimistic assumptions on the fidelity of model inputs, to a multiple, probabilistic forecasting approach that realistically considers the inherent errors and uncertainty in the model forcing data and fundamental physics. Ensemble models of large-scale, complex systems are typically created by either combining different models (called a model-intercomparison project MIP) or by perturbing input conditions and physics of a single model (Falloon et al., 2014). The MIP approach involves taking a selection of models that differ in their system representation and evaluating forecasts for a range of scenarios. This approach is susceptible to misinterpretation if the models are not independent or if they share approximations or simplifications of certain processes. Further, it does not provide insight into individual model uncertainty or performance (Davie et al., 2013). To avoid these issues (and cognizant of the limited number of operational wave-forecasting systems available), this study considers a single model from which we create ensembles. This limits our focus to one model, but it provides a more systematic analysis of associated uncertainty and sensitivity to forcing data. It also assumes that the wave model is an appropriate simulation platform, an aspect that is assessed in greater detail in section 3.2.

This paper focuses on real-time forecasting of wave conditions at a case-study site in Monterey Bay, California and is structured as follows: The methodology section describes the approach and it includes a description of the model along with the generation of ensemble predictions. This section also details the different aggregation techniques investigated. A short description of the model construction and model set-up are provided including details on inputs and forcings to the model from a suite of real-time operational forecasting platforms. The results section describes the application of the model-aggregation technique to the Bay and the ability of the scheme to generate forecasts is assessed against the measurement data. Finally, conclusions from this research are drawn and the recommendations for future research made.

2. Motivation

Circulation and mixing in coastal regions results from the complex, non-linear interactions of waves, ocean currents, and winds. Depending on local conditions, the contribution from each can vary and a comprehensive study requires simultaneous consideration (O'Donncha et al., 2015). The objective of this study is to develop a robust system to forecast wave conditions by combining physics-based models with a machine-learning technique. Our analysis demonstrates that forcing the SWAN model with a single-point measurement of wave conditions provides marginally better performance than when forced by data from WAVEWATCH III (Tolman et al., 2009), despite the greater spatial richness of these forecast data. This illustrates a central point of modeling and operational forecasting; a forecast is not necessarily a prediction on the true state of a system but rather the best estimate based on available data. The reality is that the WAVEWATCH III data provide some information on the likely state of the system rather than being viewed as a specific deterministic forecast. On the other hand, specifying boundary conditions directly from observation data is not a defensible strategy as one loses the ability to operate in forecasting mode; i.e., the model can only be run until the time of the most recent available observation data and in many respects the model then acts as a spatial interpolation module rather than a forecasting model.

We propose a framework to integrate accurate observational data with forecast conditions, to improve predictive capabilities. To investigate the likely true solution state of the system, we consider statistical perturbations of inputs (lateral wave boundary data) to the SWAN model. This yields a set of ensemble predictions for the next 48 hours. We propose a non-invasive, model-aggregating approach that integrates these models based on a set of learned weights computed by minimizing differences between model outputs and observations at each time when measured data become available. These weights are then used to produce a single, deterministic forecast cognizant of best model performance both historically and at the most recent observation. The advantages of the proposed framework are: (1) there are no restrictions on which models may be included in the ensembles and information from deterministic physics models, stochastic models, or data-driven approaches could be readily incorporated and (2) the weights are computed using model outputs, so no modification to the model is required as would be necessary for data assimilation (DA), for example. As a comparison, a number of frameworks exist to integrate DA with existing models. Examples include the Data Assimilation Research Testbed from the National Center for Atmospheric Research (NCAR, Anderson et al., 2009), the Java-based OpenDA (van Velzen et al., 2016), and the Parallel Data Assimilation Framework (Nerger et al., 2005) developed by the Alfred Wegener Institute. These provide a variety of different DA libraries and typically interface with the forecasting model in one of two ways: in-memory coupling by directly transferring data to the assimilation library (e. g., by calling the model as a subroutine to the DA framework) or through a “black box” approach in which all interactions between the models go through input and output files (Brankart and Melet, 2010). The amount of work required to couple a DA framework is model dependent and can be reduced to a question of how difficult it is to pass data in appropriate formats to the DA system (and read back) or whether the model can be readily wrapped in external code (i. e., implement the model as a library where the DA framework can readily access and update the model state). Browne and Wilson (2015) evaluated different strategies to couple DA libraries with complex models and proposed an approach using MPI that aimed to reduce changes to source code. This paper presents an alternative approach that acts only on model outputs to update forecasts based on historical performance against observations.

3. Methodology: Physics-based wave forecasting

This effort compares SWAN wave-model predictions in Monterey Bay forced by forecasts of wave conditions, ocean currents, and wind speeds to data from three National Data Buoy Center (NDBC) buoys. The wave model is driven by wave forecasts from NOAA's NCEP, current forecasts from a Regional Ocean Modeling System (ROMS) hydrodynamic model of Monterey Bay (Patterson et al., 2012), and wind forecasts from TWC.

SWAN is a third-generation wave model that estimates wave conditions from prescribed wave information along boundary segments, ocean currents, winds, and bottom bathymetry. The spectral action balance equation describes the evolution of the wave-energy density spectrum $E(\sigma, \theta)$, over frequencies σ (as observed in a frame of reference moving with the current velocity) and propagation direction θ (the direction normal to the wave crest of each spectral component) (The SWAN Team, 2006):

$$\frac{\partial N}{\partial t} + \left(\frac{\partial c_x N}{\partial x} + \frac{\partial c_y N}{\partial y} \right) + \left(\frac{\partial c_\sigma N}{\partial \sigma} + \frac{\partial c_\theta N}{\partial \theta} \right) = \frac{S_{\text{tot}}}{\sigma}, \tag{1}$$

where N is the action density defined as the ratio of energy E to relative frequency σ ($N = E/\sigma$). The first term in the equation represents the local rate of change of action density with time while the propagation of action in geographical space is represented by the second and third terms. These terms incorporate propagation velocities c_x and c_y in the x and y directions, respectively. Depth- and current-induced refractions are represented by the fourth term, which describes the propagation velocity, c_σ, in spectral space (σ, θ). Shifting of the relative frequency due to variations in depth and current are represented with the fifth term. The source term, S_{tot}, represents the effects of wave generation, nonlinear wave-wave interaction, and dissipation (O'Brien and Ragnoli, 2014).

The preceding equation can be solved as a hyperbolic equation on the discretized spectra and propagated forward in time. However, on large domains requiring high spatial resolution, this propagation in time is computationally expensive requiring matrix inversion incurring computational costs of up to $\mathcal{O}(n^3)$ (Toman, 2010), where n is number of elements. An alternative approach is to assume quasi-stationarity in the propagation of boundary effects across the model domain. This reduces (1) to an elliptic equation that can be resolved directly using an iterative solver.

3.1. SWAN Model Set-up

Figure 1 illustrates the extents of the Monterey Bay modeling domain (64×54-km^2 domain discretized across 710×480 computational elements providing a horizontal resolution of 0.001° each approximately equal to 90×110 m^2) for the SWAN model, which was originally developed by Chang et al. (2016). NOAA NDBC buoys from stations 46042 (white), 46114 (red), and 46240 (green) provide measurements of wave conditions together with other ocean and meteorological data reported every 30 to 60 minutes. Table 1 provides further information on the buoy datasets.

Primary inputs to the SWAN model are lateral boundary information of wave forecasts, along with spatial distribution of ocean-current forecasts, and wind forecasts. Lateral boundary information of wave height, direction, and period are prescribed on the Southern, Western, and Northern boundaries with data extracted from the WAVEWATCH III Eastern North Pacific model at 0.25° spacing (turquoise markers denote where WAVEWATCH III data are available and red segments where prescribed as lateral boundary conditions). Ocean currents are from the Monterey Bay ocean forecasting system (black symbols) based on the 3000-m-resolution ROMS model (IOOS, 2017). Wind speeds were extracted at 0.25° spacing (turquoise symbols) from TWC application programming interface (API). TWC provides information on a variety

Table 1: NDBC buoys used in this study.

Station name	Longitude	Latitude	Water depth (m)	First data
Buoy 46042	122.452° W	36.791° N	2,098	June 1987
Buoy 46114	122.351° W	36.723° N	1,463	September 2011
Buoy 46240	122.907° W	36.626° N	17.8	January 2009

of meteorological conditions, forecasts, alerts, and historical data, which can be extracted either directly from the TWC API or through the IBM Bluemix platform. Hourly forecast data out to fifteen days are available along with historical cleansed (i. e., subject to quality assurance procedures) data for the past 30 years. Table 2 summarizes details on the datasets used to force the SWAN model.

3.2. Verification of the SWAN Model

The first step of the study was performance confirmation of the SWAN model and assessment of the sensitivity to boundary forcing data. Six days of

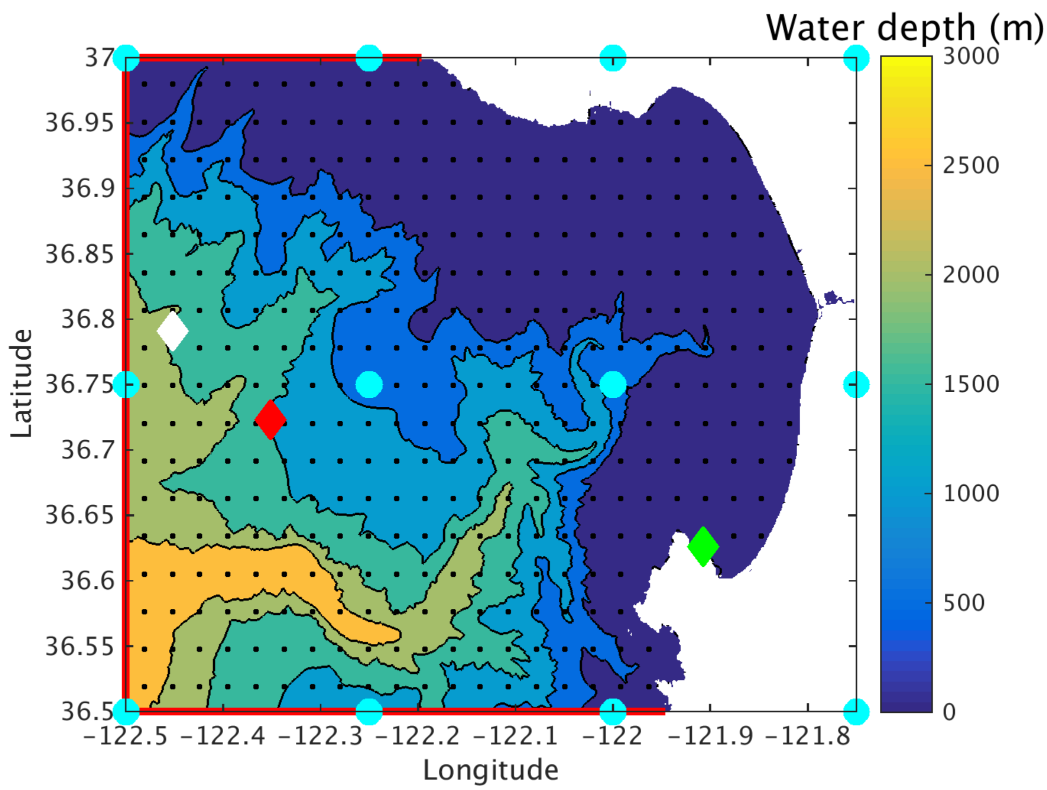

Figure 1: SWAN model domain with color indicating the bathymetric depth. The three buoys used to verify the model are indicated with the symbols where the white diamond is Buoy 46042, the red diamond is Buoy 46114, and the green diamond is Buoy 46240. The red boundaries are where wave-condition data were prescribed. The turquoise symbols indicate where WaveWatch III data were available and TWC data were prescribed to the model. Naturally WaveWatch III data returned no data over land. The black dots are where ROMS current data were provided.

NOAA wave data, ROMS ocean currents, and TWC winds were assembled into steady-state SWAN model runs at three-hour intervals. SWAN models were run with different lateral wave boundary conditions. Currents were supplied by the ROMS 3000-m grid, wind forcings were supplied by TWC, and boundary wave conditions were extracted from either WAVEWATCH III forecasts or from NOAA Buoy 46042. Figure 2 compares observed wave-characteristics data (significant wave height, H_s, top row; wave period, T, middle row; wave direction, D, bottom row) for the three buoys in the Monterey Bay area (black curves) to the different model forecasts.

The black symbols denote data extracted from WAVEWATCH III forecasts at the nearest grid point to each buoy. Given the extent of this relatively coarse model (0.25° resolution for the entire Eastern North Pacific), it is not surprising that there is significant mismatch. Moreover, the location of NOAA Buoy 46240 (green diamond in Figure 1), which is sheltered from incoming westward waves, results in a significant discrepancy between WAVEWATCH III-simulated and buoy-measured wave characteristics because the nearest WAVEWATCH III grid point is well to the southwest of the buoy (−122° W and 36.5° N). The blue symbols are simulated wave characteristics when NOAA wave data from Buoy 46042 were supplied to the SWAN model as boundary conditions. Unsurprisingly, blue and black symbols are quite similar at Buoy 46042 because this buoy served as boundary conditions to the SWAN model with results nearly as close at Buoy 46114 (red diamond in Figure 1), which is about 13 km to the east-southeast. The red symbols are simulated wave characteristics at the three buoys when the SWAN model was forced by WAVEWATCH III wave conditions along its boundaries.

We use both root-mean-squared errors (RMSE) and mean absolute per-

Table 2: Forecast data descriptions including resolutions at which they were applied to the SWAN model.

Model	Data	Resolution	Forecast (days)	Frequency (hr)
WaveWatch III	H_s, T, D	0.25°	7.5	3
TWC	u, v winds	0.25°	15	1
ROMS	u, v currents	3000 m	2	3

centage error (MAPE) to quantify model deviation from measured data

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{n=1}^{\mathcal{N}} \left(X_{\mathrm{model}} - X_{\mathrm{obs}}\right)^2}{\mathcal{N}}}, \tag{2}$$

$$\mathrm{MAPE} = \frac{100}{\mathcal{N}} \sum_{n=1}^{\mathcal{N}} \left|X_{\mathrm{model}} - X_{\mathrm{obs}}\right|, \tag{3}$$

where X_{model} and X_{obs} represent the model-predicted and observed values,

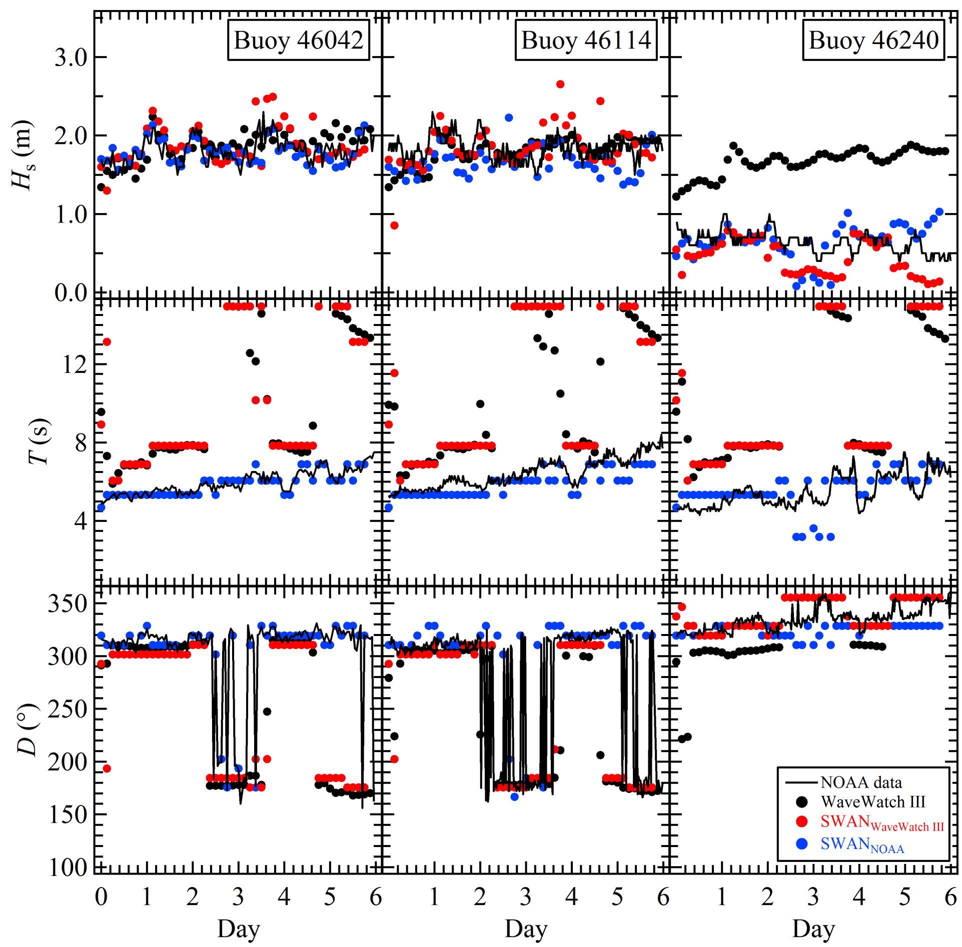

Figure 2: Comparison of simulated and measured wave characteristics at the three NOAA buoys.

respectively. As an initial assessment of model performance over the study period, RMSEs for significant wave heights at each buoy are listed in Table 3. SWAN provides slightly more accurate results using measured wave data from NOAA than when WAVEWATCH III boundary conditions were used. This comparison ignores the fact that forcing with NOAA data uses relatively accurate observation data, while forcing with WAVEWATCH III data incurs the penalty of uncertainty inherent in any operational forecasting platform. This is however balanced by the restrictive constraint that forcing with measured data does not enable forecasting. Notably, SWAN performs better at buoy 46240 when forced by WAVEWATCH III data, a location more sensitive to complex directional influences that is likely better resolved from the greater spatial coverage of WAVEWATCH III data than from a single buoy. Hence, the remainder of the study focused on ensemble generation using forecasts from the NOAA WAVEWATCH III model as forcing data. This model configuration produced mean RMSE across the three buoys of 0.57m (Table 3) corresponding to a MAPE of 22.1%. Bidlot et al. (2002) note that 40- to 60-cm RMSEs are typical for modeling studies of wave heights demonstrating that the SWAN model and configuration are appropriate as a forecasting tool.

A notable feature of Table 3 is the relatively high RMSE reported for wave direction. This is partly a result of the higher variance and volatility in the observed wave directions than simulated. Focusing on the outer buoys (46042 and 46114 in Figure 2), wave direction typically alternates between 175 and 310 degrees (where direction is denoted as direction from which waves are coming with units of degrees North, increasing clockwise - with 0° being North), denoting waves from South or from West-Northwest. These general directional trends are captured well by the model. However, periods

Table 3: RMSEs computed for the different models and model configurations against measured data from NDBC buoys. The first data row compares predictions from the WaveWatch III model against the three buoy while the next two rows present RMSEs computed for the SWAN model forced by WaveWatch III data and measured data extracted from NOAA Buoy 46042, respectively.

	Buoy 46042			Buoy 46114			Buoy 46240		
Model	H_s (m)	T (s)	D (°)	H_s (m)	T (s)	D (°)	H_s (m)	T (s)	D (°)
WAVEWATCH III	0.67	3.61	80.3	0.62	3.29	74.5	1.62	4.08	103.7
$\text{SWAN}_{\text{WAVEWATCH III}}$	0.68	3.56	80.0	0.68	3.23	73.7	0.35	4.32	16.5
$\text{SWAN}_{\text{NOAA}}$	0.50	0.57	62.3	0.41	0.45	73.9	0.42	1.06	12.0

with very high volatility in wave direction, as seen at day 3 in Figure 2, are not replicated by the model, instead predicting relatively stationary incoming waves from the South. It is likely that the higher spatial resolution (particularly of the coarse-resolution WAVEWATCH III data serving as forcing) do not encapsulate these small-scale, localized changes in wave characteristics.

4. Methodology: Integrating forecasts with machine learning

4.1. Creation of Model Ensembles

A stationary wave model is insensitive to initial conditions; instead an iterative sweep of wave conditions is conducted until the solution converges to within some threshold. Hence, we only consider perturbations to forcing conditions and model physics. Chen et al. (2004) considered the sensitivity of a global WAVEWATCH III model driven by wind forcing and updated by information derived from DA. DA provided improved short-term forecasts (12 to 24 hours) while perturbation of wind forcing had the greatest impact on forecasts. However, in regional models the sensitivity to wind forcing is less and the greatest source of uncertainty is the lateral boundary information on wave conditions specified from an external model or other data source. Preliminary studies investigating the sensitivity of the Monterey Bay SWAN model to perturbed inputs of wind forcing extracted from the NOAA Global Ensemble Forecast System (GEFS) demonstrated low sensitivity to perturbation of wind inputs (Zhang et al., 2017). As a result of the limited spatial scale of the Monterey Bay domain ($\approx$3,500 km^2), perturbing wind input data based on outputs from NOAA GEFS forecasts yielded changes in wave height of less than 0.5 cm. Considering the lack of sensitivity, and to simplify the ensemble generation process, we focused on: (1) perturbing wave boundary data prescribed on the lateral boundaries and (2) varying model physics by creating SWAN ensemble members using first-, second-, and third-generation physics (The SWAN Team, 2006).

Creation of perturbed wave boundary condition data considered the deterministic forecast from the WAVEWATCH III model and the information on system dynamics provided by the NOAA buoy nearest to the SWAN boundary, and consequently most exposed to open-ocean wave conditions (Buoy 46042, see Figure 1). To quantify the dynamics of the system, we used the standard deviation of the observation data to generate upper and lower statistical bounds on the prescribed boundary conditions that encompassed all the observation data at this buoy with a 95% confidence interval.

The buoy data were first pre-processed to remove outliers by eliminating the upper and lower 2.5% of observations. Note that we experimented with a range of percentages between 0.5% and 10% for pre-processing. The estimated standard deviation stabilized at 2.5%, which led to our choice of this threshold. Next, the standard deviation of observations was computed using a 48-hour moving window. Upper and lower bounds to the wave boundary information from the WAVEWATCH III model were computed by adding ± 1.96 standard deviations to the deterministic WAVEWATCH III forecasts. Figure 3 presents the deterministic predictions of wave heights extracted from WAVEWATCH III together with the computed statistical bounds (gray envelope) while the blue circles denote measured data. The plot illustrates that the deterministic forecast at times overestimated (e.g., July 17$^{\text{th}}$) or underestimated (e.g., July 29$^{\text{th}}$) actual observations. These temporal biases and general spread of measurements around the model predictions are encompassed by the calculated statistical bounds indicating that the generated spread encompassed most system dynamics.

We generated 12 unique ensemble members by perturbing wave boundary conditions within the statistical bounds estimated as described above. Latin hypercube sampling (LHS) was used to extract samples from within these bounds and prescribed as model boundary conditions. LHS is a statistical method for generating a near-random sample of parameter values from a distribution (McKay et al., 2000). When sampling a function, its range is divided into $\mathcal{N}$ equally probable intervals and a random sample is selected from each interval. This ensures adequate coverage of a distribution where the tails are important.

Additional ensembles were created by manipulating model physics. SWAN can operate in first-, second-, and third-generation modes. The first- and second-generation modes are essentially those of Holthuijsen et al. (1988); first-generation with a constant Phillips "constant" of 0.0081 and second-generation with a variable Phillips "constant." Third-generation processes include wind input, whitecapping, bottom friction, depth-induced wave breaking, dissipation due to vegetation, mud, or turbulence, obstacle transmission, nonlinear wave-wave interactions (quadruplets and triads) and wave-induced set-up. An overview of these options is given in SWAN User's Manual (The SWAN Team, 2006). By perturbing inputs and model physics a total of 15 unique ensemble members were generated (12 with perturbed wave boundary data using third-generation physics, and individual ensemble elements with unperturbed inputs and first-, second- and third-generation physics modes).

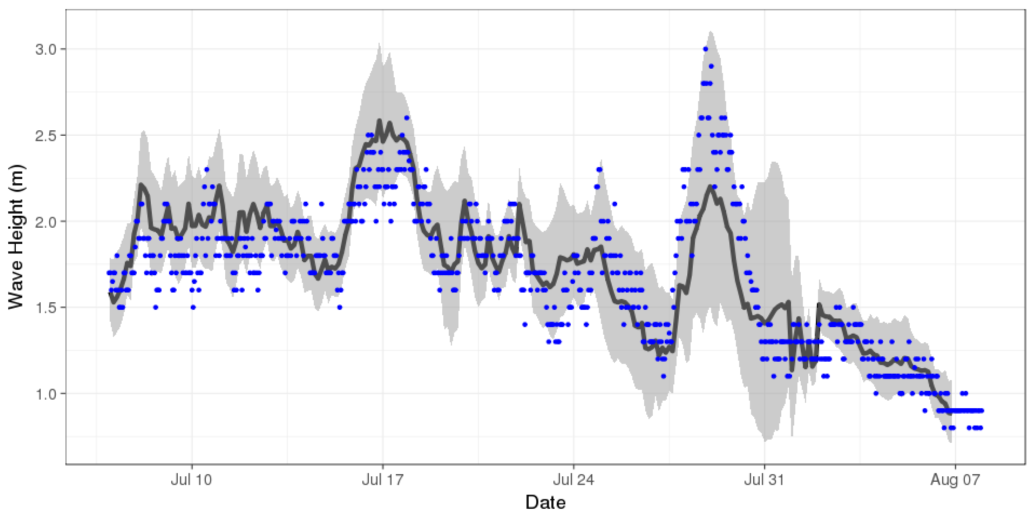

Figure 3: Time series of the mean Hs computed by WaveWatch III at the western model boundary (black curve). The blue circles denote observations from Buoy 46042 (see Figure 1) while the gray band represents the upper and lower bounds on wave height for specifying model lateral boundary conditions.

4.2. Model Aggregation

The underlying assumption of ensemble modeling is that each model contains some information pertinent to the true state of the system. The interplay between models is expected to vary in both space and time; i. e., individual models of an ensemble perform better at different points in space and time depending upon ambient conditions, individual model forcings, and other physical interactions. The objective of the aggregation method is to develop a weight for each member of the ensemble taking into account previous predictions and observations.

We considered two aggregating approaches to evaluate different sensitivities of forecasts to previous model performance. The objective is to generate a weight vector, $\mathbf{u}_t$, at each time index, t, that minimizes average mean square error (MSE) between predictions and observations by aggregating the $\mathcal{N}$ ensemble predictions into a single "best-estimate" forecast.

The first technique investigated was a ridge regression (RR) prediction algorithm. The weight vector for each time update was computed as (Mallet et al., 2009):

$$\mathbf{u}_t = \arg\min_{\mathbf{u}\in R^n} \left[\lambda ||\mathbf{u}||_2^2 + \sum_{t'=1}^{t-1} \sum_{s\in S_{t'}} \left(\mathbf{u}\cdot\mathbf{x}_{t'}^s - \mathbf{y}_{t'}^s\right)^2 \right], \tag{4}$$

where $\mathcal{N}$ is the number of ensemble members, $\mathbf{x}_{t'}$ is a vector of dimension $\mathcal{N}$ containing the prediction from each ensemble member at time t' (for each station s), $\mathbf{y}_{t'}$ represents observations at time t', $\mathbf{u}_t$ is a vector of weights computed for each ensemble member, and $S_{t'}$ represents the number of observation stations for which data are available at each time. Conceptually, the objective can be considered as choosing the appropriate weight for each ensemble element to minimize the MSE across all observation stations or buoys. The training of the weights vector on data progresses on a certain subset of data from time t' to $t-1$ from which predictions are then made for the next time step t based on the most recent ensemble predictions. λ is a penalty function that keeps the magnitude of $\mathbf{u}_t$ small and reduces variation between consecutive values. Mallet et al. (2009) provided a brief discussion of this penalization function, which is typically selected in an *ad hoc* manner for each study to balance contributions from the most recent model-observation datasets and historical data. Scope exists however to use more robust cross-validation approaches to guide parameter selection.

The computed weights from time t' to $t-1$ are then used to make a forecast, $\hat{x}_t^s$, for each station, s, at time t as:

$$\hat{x}_t^s = \mathbf{u}_t \cdot \mathbf{x}_t^s = \sum_{m=1}^{\mathcal{N}} u_{m,t} x_{m,t}^s, \tag{5}$$

where $\mathbf{x}_t^s$ is each member of the ensemble prediction at station s, and $\mathbf{u}_t$ is the weight applied to each prediction.

The second technique implemented here is an Exponentiated Gradient (EG) algorithm for linear predictors (Kivinen and Warmuth, 1997). The EG algorithm also has a weight vector $\mathbf{u}_t$ and predicts with $\hat{x}_t^s = \mathbf{u}_t \cdot \mathbf{x}_t^s$. The update of the model weights vector for each model ensemble member $x_{m,t'}$ is of the form (Kivinen and Warmuth, 1997):

$$u_{m,t} = \frac{\sum_{t'=1}^{t-1} r_{m,t'} u_{m,t'}}{\sum\limits_{j=1}^{\mathcal{N}} \sum_{t'=1}^{t-1} r_{j,t'} u_{j,t'}}, \tag{6}$$

for all $m = 1, ..., \mathcal{N}$ where t' denotes data at the previous time step and, $r_{m,t'}$ is computed as:

$$r_{m,t'} = \exp\left[\sum_{s \in S_{t'}} -2\mu(u_{m,t'} \cdot x^s_{m,t'} - y^s_{t'})x_{m,t'}\right], \tag{7}$$

where μ denotes learning rate.

Weights computed with the EG approach were normalized by the sum of all weights as expressed in (6). This constrained the weights to a convex combination where all weights summed to one as opposed to the unconstrained weights provided by RR (i. e., where weights could take any values that satisfy the objective function). A potential advantage of EG-type approaches over RR is this constraint on weights, which limits rapid fluctuations over time. Convex-combination weight vectors may be more extensible to other regions of the model domain away from where observations are available (and consequently are included in the weight computations) than unconstrained weights (Mallet et al., 2009). These aggregated predictions will always fall in the envelope of the ensemble predictions, which avoids unrealistic model forecasting.

A framework to implement the above aggregation consists of the following steps:

1. Initialize the weight vector $\mathbf{u_0} = 0$
2. Select penalization coefficient $\lambda > 0$
3. For each timestep $t' = 1...t$
   Predict with $\mathbf{u_t}$
   Using the prediction and the observations compute updated $\mathbf{u_t}$ that minimizes (4)
   Using the updated weight, $\mathbf{u}_t$, compute forecast, $\hat{x}^s_t$ for each station, s at time t using (5).

The Python toolkit SciKit-Learn (Pedregosa et al., 2011) provides a high-level programming interface to linear model implementations such as ridge regression.

5. Results and Discussion

Analysis of results and performance of the aggregation technique focused on one month of simulations from July 7$^{\text{th}}$–August 7$^{\text{th}}$, 2017. The SWAN model was set up as described in Section 3.1. Forcing data for wave conditions and wind and current speeds were extracted from the sources described in Table 2. The model ensembles comprised 15 individual elements composed of: a forecast with deterministic, one member with unperturbed wave heights, 12 simulations with perturbed wave boundary heights, and two simulations based on first- and second-generation SWAN model physics (all other ensembles used third-generation physics).

Model aggregation focused on making on-line forecasting using all available past data (real-time) to update weight coefficient applied to each ensemble element. Weights were initialized to zero, and the values for each model forecast time were computed based on a minimization of the differences between forecasts and observations together with historical information contained in the first term on the right-hand-side of (4). The fundamental ob-

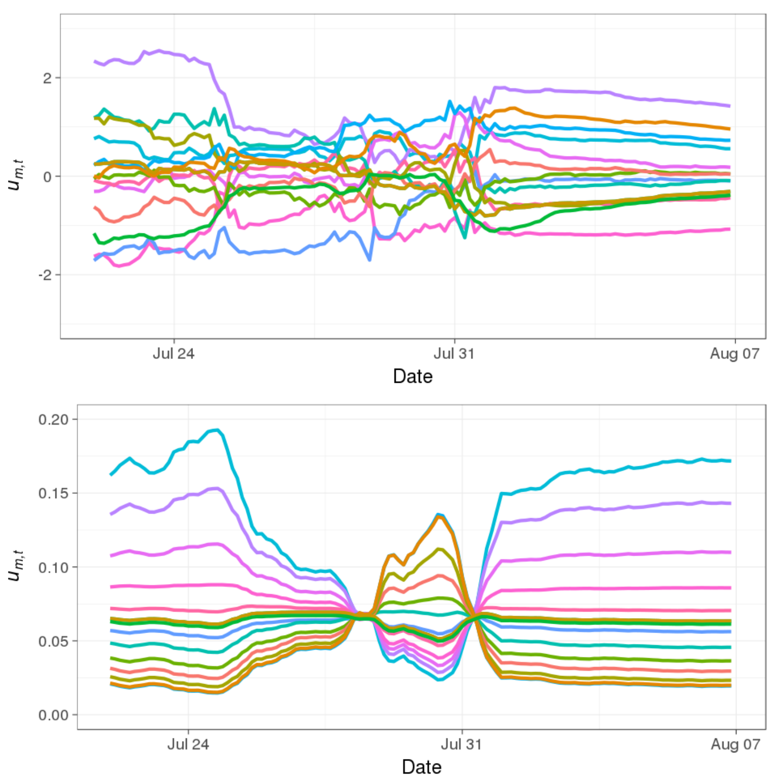

Figure 4: (a) RR and (b) EG weights computed for each of the 15-member model ensemble for each time snapshot (predictions made at three-hour intervals). Each curve represents the weight attached to an individual model and evolution over time.

jective is to leverage the observation data available in real-time to improve short-term forecast capabilities. Figure 4 presents the computed weights using the RR and EG aggregation method for a one-month period. Most models contribute to the aggregated forecast with the majority of weights having strongly non-zero values. Further, the dynamics and variations of the weighting aggregation over time were apparent with larger magnitude weights corresponding to periods with largest spread in model forecasts (and consequently highest uncertainty from the ensemble prediction perspective). During periods of large model spread, the minimization of model-observation differences was achieved by applying large weights to models that performed well and low weights to models that performed poorly.

A key consideration for wave forecasting is the temporal dynamics of the system. The fundamental aspect of weighted model aggregation is that there is a certain relationship between successive forecasts and observations; i.e., there is a likelihood that the ensemble element that performs best at time t will be the model that performs best at time $t + 1$. By constantly updating weights based on the difference between the latest observations and forecast, one can ensure that the best-performing model is assigned the highest weight. This "follow-the-leader" forecasting system works best if the quantity being modeled is relatively stationary with pronounced historic influence. Figure 5a presents the spatially averaged H_s computed at each of the three buoy locations plotted against spatially averaged observations (over the three NOAA buoys). The dashed curves represent H_s computed by each ensemble element while the solid black curve represents observation data and the solid red curve represents the model that yielded the best performance against observed data. The highly dynamic nature of the system is clear with significant variation in H_s over short time periods. Further, at various time periods, the models converged to a similar solution (e.g., August 3$^{\mathrm{rd}}$ – 7$^{\mathrm{th}}$), while at other times they spanned a broad range (e.g., July 31$^{\mathrm{st}}$). This illustrates the distinct challenges of ensemble aggregation focused on computing appropriate weights incorporating these highly dynamic events.

Figure 5b presents the evolution of MAPE for the corresponding period for three different cases: (1) where the best-performing ensemble element was considered (best individual model indicated by the black curve), (2) a weighted aggregation using the RR method with weights computed for every three-hour prediction (red curve), and (3) weighted aggregation using the EG method (blue curve). For the one-month study period, the best individual model reported a MAPE of 20.8%, while the weighted aggregation

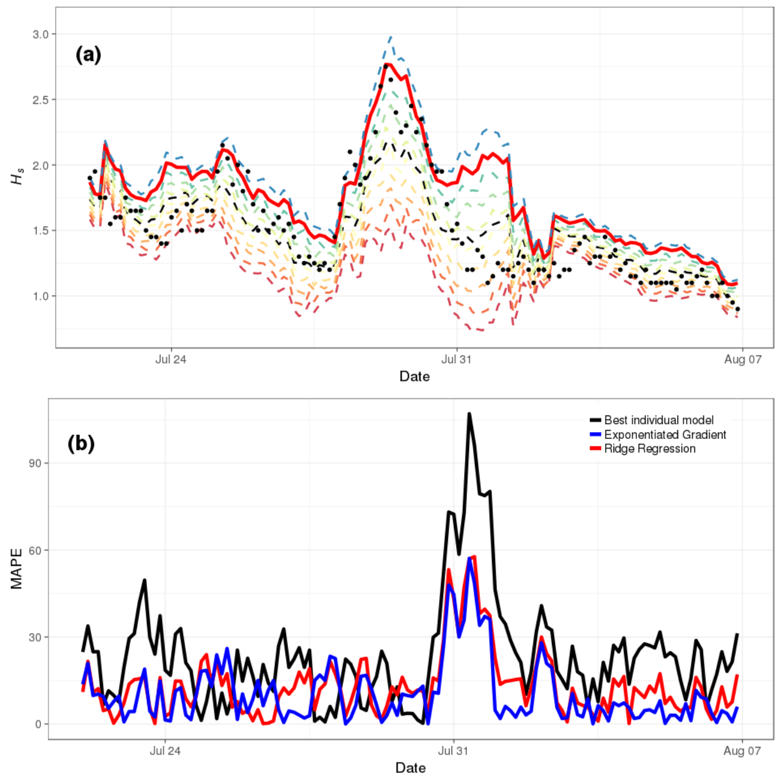

Figure 5: (a) Predicted H_{s} against observations spatially averaged across all three buoys. The solid black curve denotes observations, dashed curves represent predictions from each ensemble element, while the solid red curve denotes best-performing individual model. (b) MAPE computed against spatially averaged observations for each of (1) best-individual model (black curve), (2) aggregation using RR (red curve), and (3) aggregation based on EG (blue curve).

technique reduced MAPE to 9.76% and 9.26% for the RR and EG forecasters, respectively. The aggregated prediction significantly outperformed the best individual model aside from a short period around July 25$^{\mathrm{th}}$ and 28$^{\mathrm{th}}$ when the error of the best individual model against observations reduced to almost zero.

Ocean waves are a volatile natural process subject to rapid variation in space and time. This means that the correlation in time is relatively low and the "history effect" of prediction is not guaranteed; i. e., often the best-performing model at time t will not be the best performing model three hours later. However, these results demonstrated that most of the time the aggregated ensemble prediction provided superior results over simply selecting the best-performing individual model. The high volatility of wave conditions means that an aggregating algorithm that adjusts more quickly to the latest conditions provides better performance overall. Chen et al. (2004) observed that a wave-forecasting system updated with DA provided improved forecasts in the near-term (12 to 24 hours). However, due to the nature of such volatile systems, gains provided by DA diminished for longer forecasting windows implying that autocorrelation of the system is limited to 12 to 24 hours. Model aggregation is subject to similar trends. The weights computed based on all past model and observation data inform the best-performing model at the present time. Depending on the temporal volatility, this information is valid for a certain time window, beyond which system dynamics may evolve toward a different set of weights (i. e., ensemble element scores). The study by Chen et al. (2004) and the results presented here suggest that the latest information influences future states for short-term predictions.

Scope exists to apply machine-learning techniques to directly forecast ocean-wave conditions (James et al., 2018). A common shortcoming of data-driven techniques, however, is that while they may be accurate on average, they may miss extreme events. Integrating machine-learning approaches with physics models affords opportunities to leverage available observation data to update predictions, while also maintaining the ability to resolve extreme events as encapsulated by the fundamental equations describing the system. Weights are computed by the machine-learning approach, but the predictions themselves are made by a physical model. Mallet et al. (2009) discussed the ability of ensemble-aggregation techniques to predict extreme events applied to forecasts of ozone concentrations. They demonstrated that for the two most extreme frequency bins (low and high ozone concentrations), the aggregated predictions always outperformed the best performing individual model. Similarly, in this study the largest reductions in MAPE were achieved when wave heights were large and the best-performing model did not adequately encapsulate the data (i.e., July 30$^{\text{th}}$ – 31$^{\text{st}}$ in Figure 5a).

Another aspect worth considering is the individual models composing the ensemble prediction. In this study, a relatively simple experimental design was adopted to build the ensembles considering the statistical deviation of observation data to determine perturbations around the mean. These did not consider in detail any prior knowledge about the model (e.g., biases or efficiency at forecasting particular events or time periods, etc.). Further, due to the relatively high computational cost of the model, there were limits to the number of ensemble elements that could be created. Despite the significant improvement in accuracy provided by the aggregation techniques, the method could be further complemented by a more extensive design of the simulation exercise. In particular, the addition of other wave-forecasting models to the ensemble would be useful to counter any potential biases of the SWAN model.

6. Conclusions

In this paper, we detailed the creation and generation of ensemble predictions from statistical perturbation of forcing data. RR and EG aggregating algorithms computed deterministic forecasts from the ensemble that leveraged past observations and past model performance of each ensemble member. Results demonstrated that the aggregating forecaster significantly improved forecasts compared to the traditional civil-engineering approach based on a calibrated model or the best individual member of the ensemble to make forecasts. The aggregating algorithm reduced MAPE from 20.8% (best individual model) to 9.26% using the EG forecaster. Computed weights demonstrate that the approach leveraged most members of the ensemble to aggregate the final forecast with most computed weights having magnitudes notably different from zero.

One of the primary advantages of this approach is that it provides a non-invasive method to leverage data to improve forecasts. As the algorithm only acts on outputs from the model to compute weighted-sum predictions, it does not require any amendment or development of the source code as is necessary with traditional DA approaches. Further, the algorithm can be readily replaced with alternative local-minima approaches that better reflect the needs of a particular study (e. g., gradient-descent approaches, etc.).

Ensemble-based forecasting is a widely used technique to account for uncertainty inherent in numerical modeling studies. Leveraging multiple simulations with perturbed inputs and physics facilitates an expanded exploration of likely future conditions and provides probabilistic information on forecasts. Many decision processes however, require a single, deterministic forecast. This is typically done with some form of averaging across all ensemble members or selection of the best individual model (based on some metric). The approach presented here provides a comprehensive technique that leverages information on past model performance and observations to aggregate ensemble elements into a single forecast. The non-invasive framework can be easily integrated into an on-line operational forecasting system. This can be readily extended to other models and in particular to combining and aggregating models with different levels of complexity and different fundamental physics (e. g., combining rule-based models with data-driven models or deterministic approaches with stochastic). Future work will investigate aggregating techniques when prior information on model complexity and physics can be used to provide prior information to the aggregating technique and parameterization.

Acknowledgements

Elements of this research has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 773330.

References

Anderson, J., Hoar, T., Raeder, K., Liu, H., Collins, N., Torn, R., Avellano, A., 2009. The data assimilation research testbed: A community facility. Bulletin of the American Meteorological Society 90 (9), 1283–1296.

Bidlot, J.-R., Holmes, D. J., Wittmann, P. A., Lalbeharry, R., Chen, H. S., 2002. Intercomparison of the performance of operational ocean wave forecasting systems with buoy data. Weather and Forecasting 17 (2), 287–310.

Brankart, J.-M., Melet, A. ., 2010. Multigrid ocean data assimilation: Sesam with agrif. Tech. rep., LEGI/MEOM.

Browne, P. A., Wilson, S., 2015. A simple method for integrating a complex model into an ensemble data assimilation system using mpi. Environmental Modelling & Software 68, 122–128.

Chang, G., Ruehl, K., Jones, C., Roberts, J., Chartrand, C., 2016. Numerical modeling of the effects of wave energy converter characteristics on nearshore wave conditions. Renewable Energy 89, 636–648.

Chen, H. S., 2006. Ensemble prediction of ocean waves at NCEP. In: Proceedings of the 28^{th} Ocean Engineering Conference.

Chen, H. S., Behringer, D., Burroughs, L., Tolman, H., 2004. A variational wave height data assimilation system for ncep operational wave models. Tech. Proced. Bull., Ser. no. MMAB/2004–2004, 1–13.

Davie, J., Falloon, P. D., Kahana, R., Dankers, R., Betts, R., Portmann, F. T., Wisser, D., Clark, D. B., Ito, A., Masaki, Y., et al., 2013. Comparing projections of future changes in runoff from hydrological and biome models in ISI-MIP. Earth System Dynamics 4(2), 359–374.

Falloon, P., Challinor, A., Dessai, S., Hoang, L., Johnson, J., Koehler, A.-K., 2014. Ensembles and uncertainty in climate change impacts. Frontiers in Environmental Science 2, 33.

Holthuijsen, L., Boer, S., et al., 1988. Wave forecasting for moving and stationary targets.

IOOS, 2017. Central and Northern California Ocean Observing System, CeNCOOS.
URL http://www.cencoos.org/data/models/roms/monterey

James, S. C., Zhang, Y., O'Donncha, F., 2018. A machine learning framework to forecast wave conditions. Coastal Engineering 137, 1 – 10.
URL http://www.sciencedirect.com/science/article/pii/S0378383917304969

Kivinen, J., Warmuth, M. K., 1997. Exponentiated gradient versus gradient descent for linear predictors. Information and Computation 132 (1), 1–63.

Komen, G. J., Cavaleri, L., Donelan, M., 1996. Dynamics and Modelling of Ocean Waves. Cambridge University Press.

Mallet, V., Stoltz, G., Mauricette, B., 2009. Ozone ensemble forecast with machine learning algorithms. Journal of Geophysical Research: Atmospheres 114 (D5).

McKay, M. D., Beckman, R. J., Conover, W. J., 2000. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 42 (1), 55–61.

Mei, C. C., Stiassnie, M., Yue, D. K.-P., 1989. Theory and Applications of Ocean Surface Waves: Part 1: Linear Aspects. Part 2: Nonlinear Aspects. World Scientific.

Nerger, L., Hiller, W., Schröter, J., 2005. Pdaf-the parallel data assimilation framework: experiences with kalman filtering. In: Use Of High Performance Computing In Meteorology. World Scientific, pp. 63–83.

O'Brien, N., Ragnoli, E., 2014. A hybrid grid system for a spectral wave model. In: Proceedings of the 9^{th} International Conference on Engineering Computational Technology.

O'Donncha, F., Hartnett, M., Nash, S., Ren, L., Ragnoli, E., 2015. Characterizing observed circulation patterns within a bay using HF radar and numerical model simulations. Journal of Marine Systems 142, 96–110.

Patterson, J., Thomas, J., Rosenfeld, L., Newton, J., Hazard, L., Scianna, J., Kudela, R., Mayorga, E., Cohen, C., Cook, M., et al., 2012. Addressing ocean and coastal issues at the West Coast scale through regional ocean observing system collaboration. In: OCEANS'12. MTS-IEEE, pp. 1–8.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al., 2011. SciKit-Learn: Machine learning in Python. Journal of Machine Learning Research 12 (10), 2825–2830.
URL http://scikit-learn.org/stable/.

Rogers, W. E., Wittmann, P. A., Wang, D. W., Clancy, R. M., Hsu, Y. L., 2005. Evaluations of global wave prediction at the fleet numerical meteorology and oceanography center. Weather and Forecasting 20 (5), 745–760.

The SWAN Team, 2006. SWAN scientific and technical documentation. Tech. Rep. SWAN Cycle III version 40.51, Delft University of Technology.

Tolman, H. L., et al., 2009. User manual and system documentation of wavewatch iii tm version 3.14. Technical note, MMAB Contribution 276, 220.

Toman, H. L., 2010. A note on stationary wave modeling with Wavewatch III. Tech. rep., NOAA National Centers for Environmental Prediction.

van Velzen, N., Altaf, M. U., Verlaan, M., 2016. Openda-nemo framework for ocean data assimilation. Ocean Dynamics 66 (5), 691–702.

Zhang, Y., O'Donncha, F., James, S. C., 2017. An ensemble-based approach that combines machine learning and numerical models to improve forecasts of wave conditions. In: OCEANS'17. Anchorage, Alaska, pp. 1–7.