Ensemble model aggregation using a computationally lightweight machine-learning model to forecast ocean waves

Fearghal O’Donncha\textsuperscript{a,}* , Yushan Zhang\textsuperscript{b} , Bei Chen\textsuperscript{a} , Scott C. James\textsuperscript{c}

\textsuperscript{a}IBM Research - Ireland, Dublin
\textsuperscript{b}University of Notre Dame
\textsuperscript{c}Baylor University

Abstract

This study investigated an approach to improve the accuracy of computationally lightweight surrogate models by updating forecasts based on historical accuracy relative to sparse observation data. Using a lightweight, ocean-wave forecasting model, we created a large number of model ensembles, with perturbed inputs, for a two-year study period. Forecasts were aggregated using a machine-learning algorithm that combined forecasts from multiple, independent models into a single “best-estimate” prediction of the true state, based on historical performance relative to observations. The framework was applied to a case-study site in Monterey Bay, California. A learning-aggregation technique used historical observations and model forecasts to calculate a weight for each ensemble member. Weighted ensemble predictions were compared to measured wave conditions to evaluate performance against present state-of-the-art. Finally, we discuss how this framework, which integrates ensemble aggregations and surrogate models, can be used to improve forecasting systems and scientific process studies.

1. Introduction

In recent decades, science has made significant advances in enabling machines to understand language and process images for applications such as fa-

\textsuperscript{*Corresponding author

\textit{Email address: feardonn@ie.ibm.com} (Fearghal O’Donncha)
cial recognition, image classification, text translation, and autonomous driving. More generally, machine learning based approaches applied to other scientific domains, particularly the geosciences, are in a nascent stage. Some examples include applying Artificial Neural Networks to simulate earthquake-cycle activity [1], accelerating Eulerian fluid simulations using Convolutional Neural Networks [2], up-scaling air pollution forecasting using deep learning [3], and research to combine physics-based rules to guide machine-learning models of vertical temperature profiles in a lake [4].

A computationally lightweight emulator (a surrogate model for more complex modeling systems) has a number of applications and benefits. At its simplest levels it enables rapid process simulations over relatively long time scales (e.g., multi-year studies of ocean conditions). The computational advantage of surrogate models is important for systems affected by uncertainties in initial and boundary conditions such as risk-assessment studies that require iterative scenario modeling [5, 6, 7]. A wide range of analyses across different disciplines can benefit from this significant computational efficiency [8]. Examples where uncertainty is a major factor include uncertainty-based or Bayesian model calibration [9, 10, 11], global sensitivity analyses [12, 13, 14, 15, 7, 16, 17], and Monte Carlo-based uncertainty or reliability analyses [18, 19, 12].

There is a growing trend in the geosciences to combine deterministic and probabilistic forecasting to provide stakeholders with single, most likely forecasts together with confidence bounds on alternative outcomes [20]. This is particularly true for ocean waves where large uncertainties exist related to fundamental aspects such as the physics of wind input, dissipation, and non-linear interactions [21]. Here, we combine ensemble-forecasting and machine-learning techniques to: (1) investigate uncertainty from an ensemble modeling system with perturbed inputs, (2) leverage the advantages of computationally lightweight surrogate models, and (3) generate a forecast that is better than the best individual model prediction.

The authors recently developed and demonstrated a machine-learning surrogate model for a physics-based ocean-wave model. The model generated a nonlinear mapping of inputs (i.e., wave-characteristics boundary conditions and spatially variable ocean currents and wind speeds) to computed outputs (spatially variable significant wave height, \( H_s \), and characteristic wave period, \( T \)). The machine-learning model yielded enormous speedup (>five-thousand-fold) in computational time while maintaining accuracy that was within the confidence bounds of the physics-based model [22].
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, and parameterization schemes [23]. The fundamental objective of ensemble forecasting is to investigate inherent uncertainty to provide more accurate 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 errors and uncertainties in the model forcing data and fundamental governing equations. Ensemble aggregation techniques can extend from simple arithmetic averages of all models to machine-learning approaches that admit aggregate ensemble predictions based on weighted summation [24]. The learning-aggregation technique makes use of historical observations and model forecasts to produce a weight for each model. A linear, convex (i.e., where weights are constrained so they sum to unity) combination of model forecasts is performed with these weights to generate the best model forecast.

This study focused on an ensemble forecasting approach applied at a case-study site in Monterey Bay, California. Ensembles were created based on a careful analysis of model sensitivity to input data at three buoy locations. Model aggregation considered three approaches: (1) naïve model aggregation, (2) ridge-regression forecasting, and (3) forecasting using the exponentiated-gradient method. The paper presents a comprehensive framework to develop and aggregate ensemble model elements cognizant of the inherent uncertainties of inputs.

Our objective is to leverage the advantages of computationally lightweight surrogate models and non-invasive, ensemble aggregation techniques to provide the best-estimate of the system state.

The remainder of the paper is organized as follows. The methodology section describes the approach adopted; it includes a description of the model along with the generation of ensemble predictions. This section also details the two aggregation techniques investigated. A short description of model construction and set-up is provided including details on inputs and forcings from a suite of real-time operational forecasting platforms. Section 3 describes the application of the model-aggregation technique to the bay and the ability of the scheme to generate accurate forecasts that outperform the best individual forecast. Finally, conclusions from this research are drawn and recommendations for future research provided.
2. Methodology

In this section, we briefly describe the development of the surrogate ocean wave model, described in detail in [22]. Section 2.2.1 describes the creation of the model ensemble elements whereby the surrogate model is run multiple times with perturbed inputs based on an analysis of the system dynamics. Finally, we describe the aggregation techniques adopted that consider two different methods to compute weights for each ensemble model element based on historical agreement with observations.

2.1. Machine-learning Surrogate Model

One of the main challenges of applying machine learning to geosciences is the enormous volumes of data required to train the model notwithstanding the amount of sensor data that can be available, which are often spatially sparse and intermittent. However, when developing a machine-learning surrogate for a physics-based model, there is the luxury of being able to run the model as many times as necessary to develop a sufficient data set for training. The Simulating WAves Nearshore (SWAN) model is the industry-standard wave-modeling tool developed at the Delft University of Technology that computes wave fields in coastal waters forced by wave conditions on the domain boundaries, ocean currents, and winds [25].

To learn features of wave conditions, a supervised machine-learning model based on a multi-layer perceptron (MLP) network was used to compute significant wave heights, \( H_s \). During training, where the network was presented with examples of the computation it was learning (i.e., SWAN model runs), an optimization problem was solved until the output of the network’s last layer consistently approximated the training data set (in this case, \( H_s \)). Within a supervised-learning framework, inputs and outputs of the model are provided and the model learns optimized weights and biases that replicate the nonlinear relationship between inputs and outputs.

The MLP [26] model was trained at a case-study site in Monterey Bay. Figure 1 illustrates modeling domain (64 × 54-km\(^2\) discretized across 71 × 48 computational elements providing a horizontal resolution of 0.01° each approximately equal to 900 × 1,000 m\(^2\)) for the SWAN model, originally developed by Chang et al. [27]. NOAA National Data Buoy Center buoys from stations 46042 (white), 46114 (red), and 46240 (green) provided measurements of wave conditions together with other ocean and meteorological
data reported every 30 to 60 minutes. Inputs to the SWAN model comprised boundary-condition wave data extracted from Buoy 46042 (Figure 1), ocean-current data from a Regional Ocean Modeling System (ROMS) hydrodynamic model of Monterey Bay [28], and historical wind data from The Weather Company (TWC). These were assembled into machine learning input vectors, $x$, with outputs corresponding to the SWAN-simulated $H_s$ field, $y$. Design matrices were developed by completing 11,078 SWAN model runs dating back to the archived extent of ROMS currents nowcasts (from April 1st, 2013 to June 30th, 2017) and assembling $x$ and $y$ into $X$ and $Y$, respectively.

Because the goal of this effort was to develop a machine-learning framework to act as a surrogate for the SWAN model, the nonlinear function mapping inputs to the best representation of outputs, $\hat{y}$, was sought:

$$g(x; \Theta) = \hat{y}.$$ (1)

A sufficiently trained machine learning model yields a mapping matrix, $\Theta$, that acts as a surrogate for the SWAN model. This facilitated sidestepping of the SWAN model by replacing the solution of the partial differential equation with the data-driven machine-learning model composed of the vector-matrix operations encapsulated in (1).

MLP regression was used to reproduce the SWAN-generated $H_s$ field, $\hat{y}$. Fundamental performance was assessed by comparing against the spatially variable SWAN $H_s$ predictions, $y$. The $X$ and $Y$ data were randomly shuffled into two groups to form the training dataset composed of 90% of the 11,078 rows of data with the testing dataset the remaining 10%. Mapping matrix $\Theta$ was calculated using the training dataset and then applied to the testing dataset and the RMSE between test data vector, $y$, and its machine-learning estimate, $\hat{y}$, was calculated.

The machine-learning model demonstrated excellent performance against the SWAN model with RMSE $\approx 9$ cm, which was notably less than the confidence bounds of the model (about 40 to 50 cm [29]). Further, the computational time to make forecasts was reduced by a factor of 1/5,000th. Comparison against measured buoy data revealed the main limitation of surrogate modeling – a surrogate model will never outperform the model it emulates and the most well-designed surrogate can only hope to match the accuracy of the more expensive model.

5
2.2. Integrating Forecasts with Machine-learning Aggregation

2.2.1. Creation of Model Ensembles

As discussed previously, ensemble forecasting typically focuses on multiple simulations where anything from physical parameterizations, numerical discretization, or input data is perturbed. Machine learning models do not readily admit adjustments to model parameters (indeed, that contravenes the data-driven philosophy of the approach); instead ensemble construction focused on perturbation of model inputs. Previous studies investigating the sensitivity of the Monterey Bay SWAN model to perturbed inputs of wind forcing (extracted from the NOAA Global Ensemble Forecast System)
or GEFS) demonstrated low sensitivity to perturbation of wind inputs [30].

As a result of the limited spatial scale of the Monterey Bay model domain (≈ 3,500 km²), perturbing wind input data based on outputs from NOAA GEFS forecasts yielded changes in wave height of less than 0.5 cm. Considering this lack of sensitivity, and to simplify the ensemble-generation process, wave boundary data prescribed on ocean boundaries were perturbed.

Creation of perturbed wave-boundary-condition data required generating upper and lower bounds for $H_s$ based on data from Buoy 46042 (closest to the western boundary). The upper and lower bounds were defined as the maximum and minimum $H_s$ recorded each day. To quantify the dynamics of the system, a Gaussian process model [31] was added to these bounds with a mean of zero and standard deviation computed from the buoy data. Figure 2 presents the upper and lower bounds of the ensemble envelope where the red

Figure 2: Time series of $H_s$ from which perturbed values were selected for boundary-data specification. The blue points are measured data, the red curves span the daily maximum/minimum, and the green curves additionally include perturbations from a Gaussian process model to incorporate variable system dynamics.
curves represent the maxima and minima and the green curves include the addition of Gaussian-process-model uncertainty to those bounds.

As a preliminary analysis step, the surrogate model was forecast with 15 ensemble $H_s$ selected from within the green bounds using a standard Latin hypercube sampling technique. This is a statistical method for generating a near-random sample of parameter values from a distribution [32]. When sampling a function, its range is divided into $N$ equally probable intervals and a random sample is selected from each interval. This ensures adequate coverage of a distribution where the tails are important.

Analysis of the ensembles indicated a lack of sensitivity to $H_s$ changes at Buoy 46240 due to its sheltered location (Figure 1). Figure 3(a) presents the spectrum of wave directions at Buoy 46114 during the study period. Naturally, this single-point measurement is only representative of wave directions along the boundaries rather than reflecting conditions at all points. To better capture representative conditions, another set of ensemble members was generated with wave directions distributed into four equally sized bins ranging from 210° to 360° based on analysis of prevailing wave directions from WAVEWATCH III model data [33]. Each wave direction was then randomly sampled from these bins to encapsulate the full spectrum of potential wave
Figure 4: Time-series representation of generated ensembles (75 individual elements) upon perturbing $H_s$ and $D$. Curves represent individual forecasts for each ensemble member and the black circles denote observations from Buoy 46114 (top) and 46240 (bottom).

conditions. A total of 75 ensemble elements were generated for the two-year study period. The lightweight nature of the model required less than 10 minutes of processor time on a laptop to generate all forecasts while comparable simulations using the SWAN model would take 183 days (on equivalent compute resources). Figure 4 compares a time series of the generated ensembles to measured data for a six-month period.

2.2.2. Model aggregation

The underlying assumption of ensemble modeling is that each member 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., member models performed 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 was to develop a weight
for each member of the ensemble taking into account previous predictions and observations.

A key consideration for model-aggregating techniques is that weights have increased dependence on more recent values than those further in the past. The objective was to generate a weight vector, $u_t$, at each time index, $t$, that minimized mean square error (MSE) between predictions and observations by aggregating the $N$ 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 [24]:

$$u_t = \arg \min_{u \in \mathbb{R}^n} \left[ \lambda ||u||_2^2 + \sum_{t'=1}^{t-1} \sum_{s \in S_{t'}} (u \cdot x_{t'}^s - y_{t'}^s)^2 \right],$$

(2)

where $x_{t'}^s$ is a vector of dimension $N$ (number of ensembles) containing the prediction from each ensemble member at time $t'$ (for each station $s$), $y_{t'}^s$ represents observations at time $t'$ and station $s$, $u_t$ is a vector of weights computed for each ensemble member, $S_{t'}$ represents the number of observation stations for which data are available at each time, and $\lambda$ serves as a regularization constant used to keep the magnitude of $u_t$ small and to reduce variation between consecutive vectors. Mallet et al. [24] described this regularization (penalty) function, which is typically selected in a bespoke manner for each study to balance contributions from the most recent model-observation datasets and historical data. As $\lambda$ tends toward zero, the regression tends toward a least-squares solution. Conceptually, the objective was to assign weights to each ensemble member that minimized the MSE across all observation stations (buoys). Training the weights vector progressed on a certain subset of data from time $t'$ to $t-1$ whereupon predictions were made for the next time step, $t$, based on the most recent ensemble predictions. Cross-validation confirmed parameter selection.

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

$$\hat{x}_t^s = u_t \cdot x_t^s = \sum_{m=1}^{N} u_{m,t} x_{m,t}^s,$$

(3)

where $x_t^s$ is each member of the ensemble prediction at station $s$, and $u_t$ is the weight applied to each prediction.
The second technique explored here was an exponentiated-gradient (EG) algorithm for linear predictors \cite{34}. The EG algorithm also has a weight vector, \( u_t \), used to predict \( \hat{x}_t^* = u_t \cdot x_t^* \). The updated weight for each ensemble member \( x_{m,t} \) was \cite{34}:

\[
    u_{m,t} = \frac{\sum_{t'=1}^{t-1} r_{m,t'} u_{m,t'} \sum_{j=1}^{N} \sum_{t'=1}^{t-1} r_{j,t'} u_{j,t'}}{\sum_{t'=1}^{t-1} r_{m,t'} u_{m,t'}}, \tag{4}
\]

for all \( m = 1, \ldots, N \), and:

\[
    r_{m,t'} = \exp \left[ \sum_{s \in S_{t'}} -2\mu (u_{t'} \cdot x_s^* - y_s^*) x_{m,t'} \right],
\]

where \( \mu \) is the learning rate.

Weights computed with the EG approach were normalized by the sum of all weights as expressed by (4). This constrained the weights to a convex combination as opposed to the unconstrained weights admitted by RR (i.e., where weights could take any values that minimize the loss function). A potential advantage of EG-type approaches over RR is this constraint on weights, which limits rapid fluctuations. Convex-combination weight vectors may better extend to other regions of the model domain away from where observations were available (and consequently were included in the weight computations) than unconstrained weights \cite{24}. These aggregated predictions will always fall within the envelope of the ensemble predictions, which avoids unrealistic model forecasting.

A further aspect of the weight computation was selection of the historical window length. A na"ive approach uses all available historical data while more sophisticated implementations acknowledge that performance in the recent past is more indicative of predictive skill. Amending (2) for RR aggregation to incorporate user-specified window lengths, \( t_w \), yielded:

\[
    u_t = \arg \min_{u \in \mathbb{R}^n} \left[ \lambda |u|^2 + \sum_{t'=t-t_w}^{t-1} \sum_{s \in S_{t'}} (u \cdot x_s^* - y_s^*)^2 \right], \tag{5}
\]

which differs from (2) only in the starting index, \( t - t_w \). Similar to selection of the learning rate, \( \mu \), and the regularization constant, \( \lambda \), a cross-validation approach was adopted to identify the optimum \( t_w \) as discussed in Section \cite{3}.
3. Results and Discussion

Analysis of the results and performance of the aggregation technique focused on the two-year period 2012–2013. Model aggregation focused on making on-line forecasting using all available past data to update weights. Weights were initialized as a zero vector at time $t = 0$ and their values at each model forecast time were computed based on a minimization of the difference between forecast and observation together with historical information contained in the first term on the right-hand-side of (2). The fundamental objective was to leverage available observation data to improve short-term forecast capabilities.

Figure 5 presents a selection of model weights computed with the RR and EG aggregation methods for a representative 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 larger spreads in model forecasts (and consequently higher uncertainties from the ensemble-prediction perspective). During periods of large model spread, minimization of model-observation differences was facilitated by applying large weights to models that performed well and low weights to models that performed poorly.

A key consideration in wave forecasting is the temporal dynamics of the system. The fundamental basis 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 performed best at time $t$ will be the model that performed best at time $t + 1$. By updating weights based on the difference between the latest observations and forecast, it was ensured that the best-performing model was assigned the highest weight. This “follow-the-leader” type forecasting system works best if the quantity being modeled is relatively stationary with pronounced historic influence.

Ocean waves are highly dynamic with significant temporal variation. Hence, a key requirement of the approach was to select the parameterization that best exploited historical performance while maintaining the ability to adjust to rapidly changing wave measurements. A series of cross-validation experiments were used to select the most appropriate value for the regularization constant/learning rate for RR and EG, respectively, and also to select optimal $t_w$. Namely, for both RR and EG, 500 experiments were conducted that varied $\lambda$, $\mu$ and $t_w$. To guide the selection process, a minimum value was specified for each to avoid overfitting. For both $\lambda$ and $\mu$, a minimum value of $1.5 \times 10^{-10}$ was specified, while a minimum value of twelve hours was specified for $t_w$. Results demonstrated that $t_w = 18$ hours provided best performance for RR while $t_w = 30$ hours proved optimal for EG. For short $t_w$, RR was relatively insensitive to $\lambda$ as only recent historical effects were incorporated. For $t_w = 6$ hours, mean average percentage error (MAPE) varied by only 0.1% for $0.05 < \lambda < 1$. The constrained nature of EG aggregation meant that longer $t_w$ yielded lower MAPEs with more sensitivity to $\mu$. MAPE varied by 15% with $\mu = 1$ achieving lowest value. Generally, results demonstrated that the high volatility of ocean-wave conditions meant that historical effects were fairly low and a short $t_w$ generally afforded the best predictive skill.

From these cross-validation studies, the hyperparameters $t_w$ and $\lambda$ that minimized MAPE were specified and forecasts were created from the aggregated model predictions. Results were compared against two benchmark forecasts:

- Best individual forecast defined as the ensemble element that provided the minimum MAPE over the course of the study period.

- Ensemble average consisting of the average of all ensemble forecasts (equivalent to applying a weight of $1/N$ to each element).
Figure 6: Predicted $H_s$ (curves) and observations (orange dots) averaged across the two buoys. The red curve denotes forecasts aggregated by RR. The dashed blue curve represents EG aggregation. The dashed green curve represents the best element from the 75 member ensemble, namely the model that had lowest MAPE while the purple curve is the arithmetic average across all model forecasts. Top figure presents predictions plotted against observations, while bottom figure presents computed MAPE for each model. More specifically, MAPE varies from 10.2% and 10.8% for RR and EG respectively while a simple arithmetic average and best individual model produce a MAPE of 11.4% and 11.2% respectively.

Figure 6 compares the spatially averaged $H_s$ computed at two buoy locations (Buoy 46240 and 46114 in Figure 1) to observations. The green curve represents the best individual model while the purple curve is the arithmetic average of component models. The red curve represents forecasts computed by RR and the blue curve from EG. Results are compared to observations denoted by orange points and demonstrated that an aggregation that took into account historical model performance through weighted averaging significantly outperformed more naïve approaches. MAPE was reduced from 11.2% for the best individual model to 10.2% for the RR method and 10.8%
for the EG method. RR outperformed EG largely due to the unconstrained weights computations, which can more quickly react to rapid changes in a highly volatile system. This reduced errors, but may yield overfitting; the constrained nature of the EG weights may be more robust and applied farther from the station with more confidence than the unconstrained RR [24].

| Buoy   | Best individual | Arithmetic average | RR  | EG  |
|--------|-----------------|--------------------|-----|-----|
| 46114  | 6.7             | 7.6                | 7.3 | 8.2 |
| 46240  | 15.5            | 15.2               | 13  | 13.5|
| Average| 11.2            | 11.4               | 10.2| 10.8|

Table 1: MAPE (%) computed for each different model scenario against individual buoys and spatially averaged across both distinct buoy locations.

Table 1 presents the computed MAPE for each model implementation against individual buoys and averaged across the two locations. It is worth noting that the weight computation selects those that provides optimal agreement against the spatially averaged buoys rather than either in isolation. Further, both buoys represent very distinct conditions, namely, open ocean conditions (buoy 46114) and sheltered, near-land conditions (buoy 46240), as demonstrated in Figure 1. The resultant aggregation aims to generate a best-estimate forecast of ocean conditions across the available buoy locations and hence may best be applied far from the buoy locations (i.e. at arbitrary locations within the bay).

These results demonstrated the viability of combining data-driven ensemble aggregation techniques with lightweight surrogate-model approaches to improve predictive skill. The nominal computational expense of the machine-learning forecasting model means there is no real practical limit to the number of ensemble elements that can be generated (generating seven-day forecast takes a fraction of a second). Training the machine learning model on simulated data from a physics-based wave model (SWAN) established an upper limit on accuracy equal to that of the SWAN model, i.e., the surrogate model is trained to reproduce SWAN predicted ocean waves and not real-world data as is generally the case in machine learning. However, by combining the surrogate models with ensemble-aggregation approaches that compute individual model weights based on predictive skill, this shortcoming can be overcome.
4. Conclusions

In this paper, we detailed the creation and generation of ensemble predictions using a lightweight, machine-learning model. Identified limitations of the surrogate model were addressed by developing ensemble-aggregation techniques to minimize MSE against available measured data. RR and EG aggregating algorithms computed deterministic forecasts from the ensemble that leveraged past observations and past performance of each ensemble model. Results demonstrated that the aggregating forecaster significantly reduced error against observations and it represents a valuable framework to integrate sparse sensor data with lightweight, data-driven surrogate models.

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 update and propagation of the model state as is necessary with traditional data-assimilation approaches (and extremely difficult to integrate with machine-learning-based models); it is also simple to implement. 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.).

A promising aspect of the approach not discussed here is the potential of using this framework to investigate long-term processes like the multi-decadal geosciences study by Arandia et al. [35]. Computationally lightweight surrogate models provide opportunity to investigate system dynamics across long time-scales much simpler than would be possible with large-scale models. Further, the accuracy of these surrogate models can be improved with ensemble aggregation methods as that presented here.

Ensemble-based forecasting is a widely used technique to account for uncertainty inherent in numerical modeling studies. Leveraging multiple simulations that encompass a wide range of potential scenarios 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 outlined a comprehensive technique that leverages information from 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 consider aggregation approaches that integrate models of different complexity and confidence metrics to provide prior information to the aggregating technique and parameterization.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme as part of the RIA GAIN project under grant agreement No. 773330.

References

[1] P. M. DeVries, T. B. Thompson, B. J. Meade, Enabling large-scale viscoelastic calculations via neural network acceleration, Geophysical Research Letters 44 (2017) 2662–2669.

[2] J. Tompson, K. Schlachter, P. Sprechmann, K. Perlin, Accelerating eulerian fluid simulation with convolutional networks, arXiv preprint arXiv:1607.03597 (2016).

[3] P. Haehnel, J. Marecek, J. Monteil, F. O’Donncha, Scaling up deep learning for PDE-based models, arXiv preprint arXiv:1810.09425 (2018).

[4] A. Karpatne, W. Watkins, J. Read, V. Kumar, Physics-guided neural networks (PGNN): An application in lake temperature modeling, arXiv preprint arXiv:1710.11431 (2017).

[5] K.-T. Fang, R. Li, A. Sudjianto, Design and modeling for computer experiments, Chapman & Hall/CRC, Boca Raton, FL, 2006.

[6] T. W. Simpson, V. Toropov, V. Balabanov, F. A. Viana, Design and analysis of computer experiments in multidisciplinary design optimization: A review of how far we have come or not, in: 12th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, volume 5, pp. 10–12.
[7] A. Marrel, N. Perot, C. Mottet, Development of a surrogate model and sensitivity analysis for spatio-temporal numerical simulators, Stochastic Environmental Research and Risk Assessment 29 (2015) 959–974.

[8] S. Razavi, B. A. Tolson, D. H. Burn, Review of surrogate modeling in water resources, Water Resources Research 48 (2012) W07401.

[9] N. Bliznyuk, D. Ruppert, C. Shoemaker, R. Regis, S. Wild, P. Mugunthan, Bayesian calibration and uncertainty analysis for computationally expensive models using optimization and radial basis function approximation, Journal of Computational and Graphical Statistics 17 (2008) 270–294.

[10] S.-T. Khu, M. G. Werner, Reduction of Monte-Carlo simulation runs for uncertainty estimation in hydrological modeling, Hydrology and Earth System Sciences Discussions 7 (2003) 680–692.

[11] X. Zhang, R. Srinivasan, M. Van Liew, Approximating SWAT model using artificial neural network and support vector machine, JAWRA Journal of the American Water Resources Association 45 (2009) 460–474.

[12] E. Borgonovo, W. Castaings, S. Tarantola, Model emulation and moment-independent sensitivity analysis: An application to environmental modelling, Environmental Modelling & Software 34 (2012) 105–115.

[13] G. Blatman, B. Sudret, Efficient computation of global sensitivity indices using sparse polynomial chaos expansions, Reliability Engineering & System Safety 95 (2010) 1216–1229.

[14] M. Ratto, A. Pagano, P. Young, State dependent parameter metamodelling and sensitivity analysis, Computer Physics Communications 177 (2007) 863–876.

[15] C. B. Storlie, H. D. Bondell, B. J. Reich, H. H. Zhang, Surface estimation, variable selection, and the nonparametric oracle property, Statistica Sinica 21 (2011) 679.
[16] V. Ciriello, V. Di Federico, M. Riva, F. Cadini, J. De Sanctis, E. Zio, A. Guadagnini, Polynomial chaos expansion for global sensitivity analysis applied to a model of radionuclide migration in a randomly heterogeneous aquifer, Stochastic Environmental Research and Risk Assessment 27 (2013) 945–954.

[17] E. Volkova, B. Iooss, F. Van Dorpe, Global sensitivity analysis for a numerical model of radionuclide migration from the RRC “Kurchatov Institute” radwaste disposal site, Stochastic Environmental Research and Risk Assessment 22 (2008) 17–31.

[18] D. Shrestha, N. Kayastha, D. Solomatine, A novel approach to parameter uncertainty analysis of hydrological models using neural networks, Hydrology and Earth System Sciences 13 (2009) 1235–1248.

[19] M. T. Schultz, M. J. Small, P. S. Fischbeck, R. S. Farrow, Evaluating response surface designs for uncertainty analysis and prescriptive applications of a large-scale water quality model, Environmental Modeling & Assessment 11 (2006) 345–359.

[20] A. E. Raftery, Use and communication of probabilistic forecasts, Statistical Analysis and Data Mining: The ASA Data Science Journal 9 (2016) 397–410.

[21] P. A. Janssen, Progress in ocean wave forecasting, Journal of Computational Physics 227 (2008) 3572–3594.

[22] S. C. James, Y. Zhang, F. O’Donncha, A machine learning framework to forecast wave conditions, Coastal Engineering 137 (2018) 1 – 10.

[23] H. S. Chen, Ensemble prediction of ocean waves at NCEP, in: Proceedings of the 28th Ocean Engineering Conference.

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

[25] The SWAN Team, SWAN Scientific and Technical Documentation, Technical Report SWAN Cycle III version 40.51, Delft University of Technology, 2006.
[26] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.

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

[28] J. Patterson, J. Thomas, L. Rosenfeld, J. Newton, L. Hazard, J. Scianena, R. Kudela, E. Mayorga, C. Cohen, M. Cook, et al., Addressing ocean and coastal issues at the west coast scale through regional ocean observing system collaboration, MTS/IEEE OCEANS’12 (2012) 1–8.

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

[30] F. O’Donncha, Y. Zhang, S. C. James, An integrated framework that combines machine learning and numerical models to improve wave-condition forecasts, Journal of Marine Systems (2018).

[31] C. E. Rasmussen, Gaussian processes in machine learning, in: Advanced Lectures on Machine Learning, Springer, 2004, pp. 63–71.

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

[33] A. Chawla, D. Spindler, H. Tolman, 30 Year Wave Hindcasts using WAVEWATCH III® with CFSR winds-Phase I, Technical Report MMAB Contribution No. 302, NOAA/NWS/NCEP/MMAB, College Park, MD, 2012.

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

[35] E. Arandia, F. ODonncha, S. McKenna, S. Tirupathi, E. Ragnoli, Surrogate modeling and risk-based analysis for solute transport simulations, Stochastic Environmental Research and Risk Assessment (2018) 1–15.