GLOBAL SHORELINE FORECASTING USING SATELLITE-DERIVED DATA AND INTERPRETABLE MACHINE LEARNING

Mahmoud Al Najar, ISAE/LEGOS (CNES/CNRS/IRD/UT3), University of Toulouse, mahmoud.al-najar@isae-supaero.fr
Rafael Almar, LEGOS (CNES/CNRS/IRD/UT3), University of Toulouse, rafael.almar@ird.fr
Grégoire Thoumyre, LEGOS (CNES/CNRS/IRD/UT3), University of Toulouse, gregoire.thoumyre@univ-tlse3.fr
Erwin W. J. Bergsma, Earth Observation Lab, The French Space Agency (CNES), erwin.bergsma@cnes.fr
Jean-Marc Delvit, Earth Observation Lab, The French Space Agency (CNES), jean-marc.delvit@cnes.fr
Dennis G. Wilson, ISAE-SUPAERO, University of Toulouse, dennis.wilson@isae-supaero.fr

ABSTRACT

Coastal development and climate change are changing the geography of our coasts, while more and more people are moving towards the coasts. Recent advances in artificial intelligence and remote sensing allow for the automatic analysis of observational data at a global scale. Symbolic Regression (SR) is a family of Machine Learning (ML) algorithms for constructing symbolic mathematical expressions which model the relations between inputs and outputs in training data. In this work, we make use of SR and a novel global-scale shoreline forecasting dataset in order to construct globally-applicable and interpretable shoreline forecasting models, and we demonstrate the potential of SR as an alternative data-driven method for coastal modelling and prediction tasks.

METHODS

This work makes use of Cartesian Genetic Programming (CGP) (Miller, 2011) in order to perform SR. CGP is a genetic programming framework that uses an acyclic graph structure in order to encode computer programs (I.e. equations, code). CGP traditionally uses a $1 + \lambda$ genetic algorithm in order to optimize the models, where the highest performing individual is selected at each generation in order to create λ offspring to act as the subsequent generation. Here, we make use of a multi-objective evolutionary algorithm (NSGA-II) (Deb et al. 2002) in order to steer the optimization procedure according to prediction performance across different coastal zones.

Furthermore, due to its white-box nature, CGP enables knowledge reuse through the direct initialization of the evolvable models with equations from existing physical models. Here, we initialize our models as ShoreFor, a common wave-driven shoreline forecasting model (Splinter et al. 2014). Initializing the algorithm with an established model has been shown to aid in algorithm convergence in previous work (Al Najar et al. 2022).

GLOBAL SHORELINES DATASET

This study makes use of a global dataset of monthly satellite-derived shorelines time series dataset (Almar et al. 2023). The dataset spans 20 years (1999-2019) and covers 8857 coastal points around the globe; in addition to global datasets of shoreline change drivers including coastal waves, sea level anomaly, and regional river discharge.

The shorelines dataset was constructed using the Google Earth Engine based on satellite image time series from the Landsat 5, 7 and 8 missions. The methodology followed to derive shoreline positions makes use of Normal Difference Water Index (threshold of 0.5) maps in order to segment satellite imagery into land and sea surfaces; the coastline position is then identified as the interface between land and sea. The methods used to used to derive the shoreline change drivers time series range from satellite altimetry (SSALTO/DUACS) to detect regional sea level anomaly (SLA), to climate reanalysis (ERA5) for wave conditions (i.e. height Hs, period Tp, and direction Dir), and land surface model simulations (ISBA-CTRIP) for river discharge.

OBJECTIVE FUNCTION DEFENITION

Three main unique drivers of shoreline change are identified in the dataset including coastal wave energy, sea level anomaly, and regional river discharge. Following Almar et al. (2023), the dominant driving force at a coastal point is determined by examining the explained variances of the different drivers at that coastal point, where a driver that explains 40% or higher of the variance at a coastal point is considered dominant. Local zones of coastal points are constructed by grouping points that are similarly-driven and where the difference (error) in their shoreline time series does not exceed 0.4 NMSE. Based on this technique, N zones (here N=3) are sampled per driver in order to form a single cluster, to be used as a unique objective for optimization. During evolution, a single coastal point is randomly sampled from each cluster in order to evaluate model performance over the different fitness dimensions (differently-driven coastal points). Using NSGA-II, the resulting populations are composed of expert models which achieve the highest prediction performances over individual optimization objectives, in addition to generalist models which provide a balance between prediction accuracy over specific objectives, in addition to accuracy across fitness dimensions.

We measure model performance using the Mielke skill test proposed by Duveiller et al. (2016). The Mielke index is a distrubution-normalized error metric and can be computed as $\lambda = 1 - \frac{N^{-1}\sum_{i=1}^{N}(o_i - m_i)^2}{\sigma_o^2 + \sigma_m^2 + (\hat{o} - \hat{m})^2}$.

The value of λ is bound similarly to Pearson correlation, where a value of 0 denotes poor model performance and a value of 1 represents perfect prediction skills.

RESULTS

In order to evaluate the method, the algorithm is run for 10,000 generations and over 25 different runs with different seeds. The resulting evolved populations are merged at the end of evolution, and generalist models are identified as the models with the highest average global correlation. We present here the performance of the highest-performing generalist model obtained from these runs and we compare its performance to the baseline ShoreFor model.

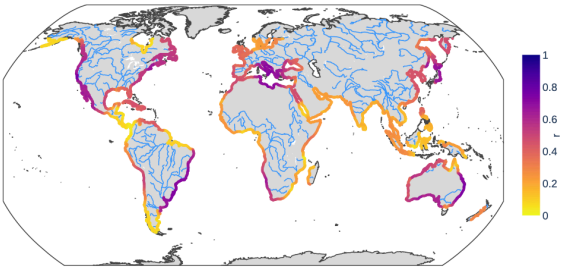

Figure 1 – The global correlation map of the evolved generalist model ($\hat{r}$ = 0.34).

Figure 1 presents the correlation map of the evolved model with the highest average global correlation. Compared to, ShoreFor, the evolved model makes use of both SLA as well as its' wave energy inputs in order to model shoreline change, and achieves a 40% increase in average global correlation (ShoreFor $\hat{r}$ = 0.24).

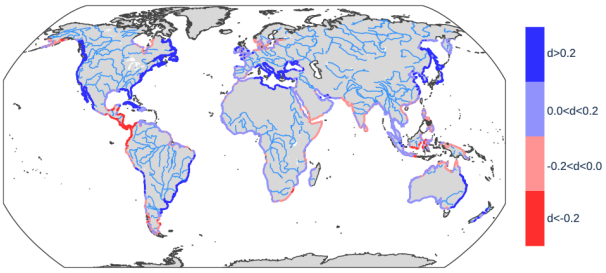

Figure 2 – The performance improvement/degradation map of the evolved model compared to ShoreFor.

Figure 2 highlights the areas where the evolved model achieves an improvement -or degradation- with respect to ShoreFor, and demonstrates a clear separation of areas where significant improvement is achieved such as North America, the eastern coast of South America, East Australia, as well as the Mediterranean Sea and the Sea of Japan; while the zone connecting the two Americas is the only area where the evolved model achieves significantly lower skills compared to ShoreFor.

While our aim is to demonstrate the ability of SR to construct compositions of interpretable used-defined functions which improve upon the performance of the baseline model, we find that further work is needed in order to ensure the physical soundness of the evolved models, thereby maximizing the potential of SR in improving both the predictive performance and the scientific understanding of shoreline change.

DISCUSSION & CONCLUSION

This short communication presents our experimental results on the use of Symbolic Regression in order to improve a physics-based shoreline forecasting model, ShoreFor. The baseline model was encoded as a CGP computational graph, and evolved using NSGA-II according to model performance over areas with different dominant drivers of shoreline change.

In order to maximize the interpretability of the resulting graphs, and consequently the potential of the method in contributing insight and understanding of the modelled system, further development of the algorithm should explore automatic graph pruning during evolution to minimize redundancy in the evolved graphs, as well as the use of dimensionality constraints in order to limit the space of possible mutations to physically-valid modifications only, through the use of grammar-guided genetic programming (Whigham, 1995) for instance.

Overall, our work demonstrates the potential of symbolic regression and CGP in evolving interpretable shoreline forecasting models which improve upon the performances of ShoreFor at a global scale, and expand its applicability to new coastal areas. We also highlight the ability of the method to combine multiple shoreline change drivers in order to model shoreline change. Finally, the results presented here motivate further study on the application of interpretable ML to open problems in coastal science due to the ability of the method to evolve models that are both well performing and interpretable.

REFERENCES

Miller (2011): Cartesian genetic programming. In Cartesian Genetic Programming (pp. 17-34). Springer, Berlin, Heidelberg.

Splinter, Turner, Davidson, Barnard, Castelle, Oltman-Shay (2014): A generalized equilibrium model for predicting daily to interannual shoreline response. Journal of Geophysical Research: Earth Surface, 119(9), 1936-1958.

Al Najar, Almar, Bergsma, Delvit, Wilson (2022): Genetic improvement of shoreline evolution forecasting models. Proceedings of the Genetic and Evolutionary Computation Conference Companion.

Almar, Boucharel, Graffin, Abessolo, Thoumyre, Papa, Ranasinghe, Montano, Bergsma, Baba, Jin (2023): Influence of El Niño on the variability of global shoreline position. Nature Communications. 2023 Jun 12;14(1):3133.

Duveiller, Fasbender, Meroni (2016): Revisiting the concept of a symmetric index of agreement for continuous datasets. Scientific reports, 6(1), 19401.

Deb, Pratap, Agarwal, Meyarivan (2002): A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE transactions on evolutionary computation, 6(2), 182-197.

Whigham (1995): Grammatically-based genetic programming. In Proceedings of the workshop on genetic programming: from theory to real-world applications (Vol. 16, No. 3, pp. 33-41).