Journal Pre-proof

Machine learning feature analysis illuminates disparity between E3SM climate models and observed climate change

J. Jake Nichol, Matthew G. Peterson, Kara J. Peterson, G. Matthew Fricke, Melanie E. Moses

PII: S0377-0427(21)00070-4
DOI: https://doi.org/10.1016/j.cam.2021.113451
Reference: CAM 113451

To appear in: *Journal of Computational and Applied Mathematics*

Received date : 1 October 2020

Please cite this article as: J.J. Nichol, M.G. Peterson, K.J. Peterson et al., Machine learning feature analysis illuminates disparity between E3SM climate models and observed climate change, *Journal of Computational and Applied Mathematics* (2021), doi: https://doi.org/10.1016/j.cam.2021.113451.

Highlights

Machine Learning Feature Analysis Illuminates Disparity Between E3SM Climate Models and Observed Climate Change

J. Jake Nichol, Matthew G. Peterson, Kara J. Peterson,
G. Matthew Fricke, Melanie E. Moses

- Machine learning models are useful for learning about observed and simulated climate data, allowing us to compare what is learned from each.
- Random forest regression models were able to learn the data and provide a means for analyzing the data's feature importance.
- Using Gini importance, we found some key differences between climate simulations (E3SM) and observed data, i.e. June and July sea ice volume and August sea ice extent are too influential on simulations.

Machine Learning Feature Analysis Illuminates Disparity Between E3SM Climate Models and Observed Climate Change

J. Jake Nichol[a,b,*], Matthew G. Peterson[a], Kara J. Peterson[a], G. Matthew Fricke[b], Melanie E. Moses[b,c]

[a] *Sandia National Laboratories, 1515 Eubank Blvd SE, Albuquerque, NM, 87123, USA*
[b] *University of New Mexico, 1 University of New Mexico, Albuquerque, NM, 87131, USA*
[c] *Santa Fe Institute, 1399 Hyde Park Rd, Santa Fe, NM, 87501, USA*

Abstract

In September of 2020, Arctic sea ice extent was the second-lowest on record. State of the art climate prediction uses Earth system models (ESMs), driven by systems of differential equations representing the laws of physics. Previously, these models have tended to underestimate Arctic sea ice loss. The issue is grave because accurate modeling is critical for economic, ecological, and geopolitical planning. We use machine learning techniques, including random forest regression and Gini importance, to show that the Energy Exascale Earth System Model (E3SM) relies too heavily on just one of the ten chosen climatological quantities to predict September sea ice averages. Furthermore, E3SM gives too much importance to six of those quantities when compared to observed data. Identifying the features that climate models incorrectly rely on should allow climatologists to improve prediction accuracy.

Keywords:
Arctic, climate, machine learning, Energy Exascale Earth System Model

*jjaken@unm.edu, Department of Computer Science, MSC01 1130, 1 University of New Mexico, Albuquerque, NM 87131, USA

Abbreviations: Energy Exascale Earth Systems Model (E3SM); Earth system model (ESM); sea ice extent (SIE); sea ice volume (SIV); total cloud cover percentage (CLT); downward longwave flux at surface (FLWS); pressure at the surface (PS); temperature at the surface (TS); near-surface specific humidity (SSH); sea surface temperature (SST); wind u component/zonal (uwind); wind v component/meridional (vwind)

1. Introduction

We have observed dramatic declines in Arctic sea ice since the advent of satellite imaging [1]. This change is of critical importance to global economic, social, political, and ecological landscapes, not least because of the opening of new navigable sea routes and the impact on wildlife [2, 3]. As an essential component of the Earth's climate, sea ice loss drives the positive feedback between surface albedo and Arctic warming and may contribute to changes in ocean circulation and mid-latitude weather [4, 5, 6, 7].

Earth system models (ESMs) provide state of the art simulations of the global climate. They include general circulation and thermodynamic models for ocean and atmosphere, and models for land, sea ice, and land ice processes. Collecting an ensemble of parameterized ESM runs produces a distribution of forecasts that provide bounds on predictions. Simulations of Arctic sea ice in these models include complex interactions between the ice, ocean, and atmosphere. However, limitations in ESMs, such as the inability to resolve critical small-scale processes, can lead to biases when compared to observations. It is, therefore, critical to identify sources of bias.

Previous generations of ESMs have, on average, underestimated the rate of sea ice loss in the Arctic [8]. This is apparent in data from the Coupled Model Intercomparison Project (CMIP), which includes simulation results from a broad array of ESMs from modeling centers around the globe. CMIP *phases* mark improvements in the state of the art. The extent of sea ice loss has been a consistent problem, first identified in phase 3 [9, 10]. By phase 5 (CMIP5), overall model bias had improved [11]. However, Rosenblum and Eisenman [8], in an analysis of 118 simulation runs from 40 CMIP5 simulations, found that 89% of CMIP5 model runs underpredicted the rate at which sea ice extent is lost (km^2/decade) by more than a standard deviation; and 2014 loss by an average of 2 million km^2. The disagreement with observation may imply that ESMs' parameters are not well-tuned. Stroeve et al. [10] suggest this discrepancy is due to missing key causal mechanisms or represent a misunderstanding of underlying physical processes.

The Energy Exascale Earth System Model (E3SM) [12], developed by the United States Department of Energy (DOE), is included in phase 6 (CMIP6) [13] (March 2019). E3SM is a new state of the science climate modeling and prediction project. In CMIP5 and E3SM, the rates of pan-

Arctic sea ice change are similar to observation before 1996 but deviate from observation afterward. In CMIP5's case, the rate of loss is less than observed [8], while E3SM's is greater than observed (Section 3.1: Data). These differences in sea ice loss rates lead to inaccurate long term predictions about absolute sea ice extent in the Arctic. To our knowledge, our work is the first mechanistic analysis of E3SM accuracy.

We use random forest regression (RFR) [14] and Gini importance [15] to determine which E3SM features drive climate predictions. We perform an identical study of historical observations to identify the features that are most influential on prediction of actual sea ice loss. By comparing the two, we determined that E3SM relies too heavily on some features, to the detriment of others, resulting in a divergence from observation. This work elucidates differences in sea ice response between observational data and E3SM simulations and can help improve sea ice prediction.

2. Related Work

Stroeve et al. [16] analyze the agreement between simulated Arctic models, CMIP3 and CMIP5, and observed data. They report that while phase 5 models are an improvement over phase 3 they consistently overestimate forecasted ice extent in the Arctic. The authors suggest that modeling may be improved by including more complex mechanisms such as sea ice albedo parameterization, thickness distributions, and melt ponds.

Rosenblum and Eisenman [8] examined CMIP5's sea ice extent predictions in the Arctic and found overprediction of sea ice extent. Correcting the models required an increase in warming well above observed rates, leading the authors to conclude that the current methods were systematically flawed.

Ionita et al. [17] presented a method for using multiple linear regression to predict the September sea ice extent minimums in the pan-Arctic region and the East Siberian Sea. Notably, they used step-wise regression because it may highlight the underlying coupled physical mechanisms between factors. For the pan-Arctic region, their model was able to predict sea ice extent anomalies for May, June, and July fairly accurately (reporting r-values between 0.84 and 0.9). Although they found a "skillful" model could be built from their list of Arctic features, they did not analyze the relative importance of those features for their models.

Reid and Tarantino used support vector regression (SVR) to predict the Arctic sea ice extent [18]. SVRs were able to construct predictive models,

but they only considered sea ice extent as a predictor and could not analyze any other features for their importance. They chose SVRs because they are successful in predicting complex dynamical systems such as climate. The authors reported the comparative results of tuning the SVR, and compared them to CMIP5 ensembles but not to observation.

3. Data and Methods

Our methods were able to account for discrepancies in climate simulations and observations. Like multiple linear regression and its associated term-weights, random forests are a machine learning method that is wholly transparent [14], unlike many other so-called "black box" methods, such as SVRs. We used RFRs and their corresponding Gini importance measure to determine how much influence each input feature has on E3SM predictions. With those tools, we analyzed each feature's impact on historical sea ice extent and used that information to highlight discrepancies with E3SM.

3.1. Data

Our machine learning (ML) models used monthly averages of June, July, and August data from the atmosphere, ocean, and sea ice to predict September sea ice extent for a given year. Results from observational and reanalysis data products are then compared against results from five ensemble members of the E3SM *historical* dataset. The features our ML models are trained on are a subset of physical quantities simulated by E3SM in the Arctic. We chose these features because they match observable features in nature and that we hypothesized would be good predictors of sea ice loss. Each feature of each dataset is a time series beginning with the start of the satellite era in 1979 and ending with the last year of available E3SM output, 2014.

The observational data included monthly sea ice extent computed from gridded, daily, passive-microwave satellite observations of sea ice concentration provided by the National Snow & Ice Data Center [19]. Sea ice concentration is a percentage value of ice in each grid cell, and sea ice extent (SIE) is computed as the total area of cells containing more than 15% ice. Sea ice volume reanalysis data were provided by the Pan-Arctic Ice Ocean Modeling and Assimilation System [20]. Atmospheric data (total cloud cover percentage (CLT), downward longwave flux at surface (FLWS), pressure at the surface (PS), near-surface specific humidity (SSH), temperature at the surface (TS), wind u component/zonal (uwind), and wind v component/meridional

Table 1: **Training Features and June Data Excerpt**: total cloud cover percentage (CLT), downward longwave flux at surface (FLWS), pressure at the surface (PS), sea ice extent (SIE), sea ice volume (SIV), near-surface specific humidity (SSH), sea surface temperature (SST), temperature at the surface (TS), wind u component/zonal (uwind), and wind v component/meridional (vwind). Values listed are means over the pan-Arctic grid for each day of the month, rounded to two-decimal places for display only.

	June										Sept.
Year	CLT (%)	FLWS (W/m^2)	PS (Pa)	SIE ($10^6 km^2$)	SIV ($10^6 km^3$)	SSH (mg/kg)	SST (°C)	TS (°C)	uwind (m/s)	uwind (m/s)	SIE ($10^6 km^2$)
1979	42.08	256.56	97 930	12.53	29.79	4.31	0.56	273.46	0.94	0.48	5.90
1980	40.89	259.51	97 901	12.20	29.15	4.44	0.68	274.67	0.99	0.47	6.83
1981	40.47	258.13	98098	12.43	26.82	4.27	0.65	274.27	0.06	0.06	6.40
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2012	40.36	271.60	98 105	10.67	16.00	5.12	1.39	277.28	−0.03	−0.06	3.55
2013	40.66	266.93	97 989	11.36	17.54	4.98	1.26	276.50	0.93	0.42	5.27
2014	39.84	263.94	98.19	11.03	17.68	4.72	1.47	275.67	0.00	0.04	5.38

(vwind)) were from an atmosphere reanalysis provided by the National Centers for Environmental Prediction [21]. Sea surface temperature (SST) was provided by the National Oceanic and Atmospheric Administration [22]. For each of the atmospheric data variables, as well as SST, monthly Arctic area averages were computed from the global gridded fields.

We used the DOE's E3SM for climate simulation data in this work [12, 23]. E3SM version 1 was a fork of the community Earth system model [24], which was a part of the CMIP5 collection analyzed by Rosenblum and Eisenman [8]. E3SM is a global model comprised of submodels for land, atmosphere, land ice, sea ice, oceans, and rivers. Specifically, we used data from E3SM's *historical* ensembles 1-5 at one-degree global resolution.

E3SM published five historical ensemble runs to offer a distribution of forecasts. The runs were initialized from different years of a 500-year pre-industrial control simulation. The historical runs start in 1850, running for 165 years to 2014. The final 36 years, 1979 to 2014, were used in our analysis to match the years of observed data. Small differences in each run's initial conditions can significantly impact long-term results, though average behavior between runs is expected to be consistent.

Table 1 summarizes the observed features we collected; an excerpt of June values is included. Each feature is a time series of the feature's mean in a given month from 1979 to 2014. Values in the time series are an area-sum over the pan-Arctic oceanic region. Each feature's monthly data is a mean

of every Arctic sample in the given month, resulting in a single value per month. Generally, the observational and reanalysis datasets have similar magnitudes to the simulation data. However, for CLT, the NCEP reanalysis is significantly lower than the E3SM data. This is a known bias in the NCEP reanalysis data, and future work could investigate feature analyses of alternative reanalysis datasets [25].

The data used in this work is publicly available on the E3SM website. The five historical ensemble runs were retrieved from the v1 one-degree data CMIP6 release. To disambiguate them from our machine learning models and observed data, we will refer to E3SM's *historical ensembles 1-5* as *simulations 1-5*, simulation runs, or simply E3SM runs for the remainder of this paper. Figure 1 shows a comparison of the observed and simulation datasets evaluated in this work.

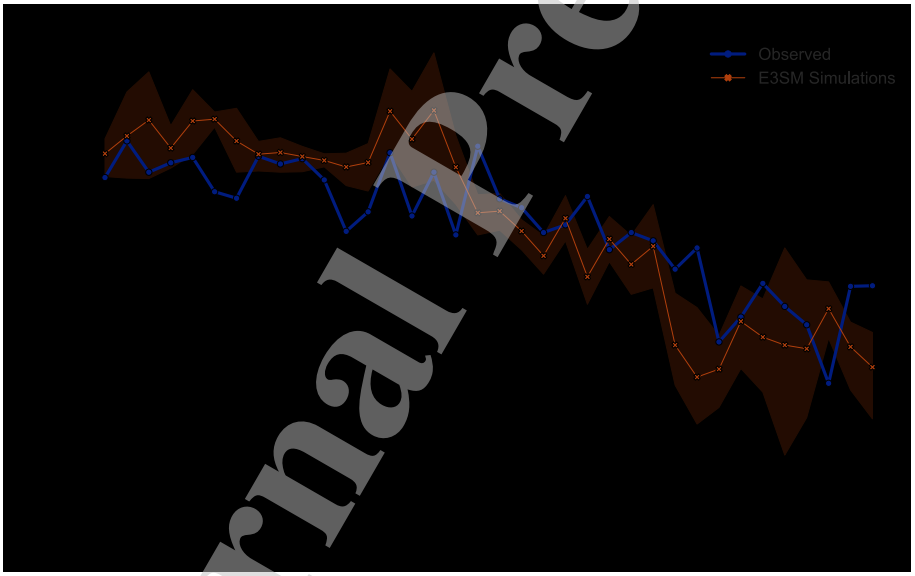

Figure 1: Comparison of observed, pan-Arctic mean September sea ice extent with predictions from E3SM's historical ensembles 1-5. The mean of E3SM simulations is shown with 95% confidence interval (shaded).

3.2. Random Forests

We found that linear models performed poorly on our data. For this work, we used RFR models because they are relatively simple, intuitive models that

can learn nonlinear relationships between features. As a part of their training, the decision trees in random forests generate Gini impurity measures. These measures are aggregated after training to determine the Gini importance of each feature. In our case, we computed importance as the total reduction in mean absolute error (MAE) caused by each feature.

RFR is an ensemble learning technique, similar to a combination of bootstrap aggregation (bagging [26]) and decision tree regression. Bagging is a method to combine the knowledge of many naive estimators, or trees in our case, by providing a subset of the full sample set to each estimator. The result is the average of many noisy, but unbiased, estimators, reducing overall variance. Random forests improve the bagging method by choosing random subsets of the feature set for each node split in every tree [27]. The number of random features each node considers, and when to split are tuned hyperparameters. The final forest's estimate is the average prediction from the random trees.

For N trees, $T_1, ..., T_N$, random forest regression prediction is computed as follows:

$$RF(N) = \frac{1}{N} \sum_{n=1}^{N} T_n(x)$$

given the training sample, x.

The random forest implementation we used was the random forest regressor from Python's sci-kit learn package [28]. The implementation uses a perturb and combine technique [29] made for tree regressors. Perturb and combine reduces test set error by introducing a diverse set of regressors via randomized regressor construction. For the rest of the data analysis, we used Python's Numpy package [30]. We utilized Python's Seaborn package [31] for data visualization.

3.3. Pre-Processing

To prepare the data for training, we split it into training and testing years. Our goal was not to develop predictive models for next year's sea ice extent. We were more interested in finding models that have learned the data well that we then used for feature analysis. Thus, we split the training and testing data randomly.

Because some years are easier to forecast than others, we should model every combination of training and testing years. For 36 total years and 18 testing years, we computed $\binom{36}{18} = 9\,075\,135\,300$ total combinations of

training and testing years. Since it is infeasible to train that many models and evaluate each feature's importance, we used this standard method to compute a sample size:

$$\frac{(z\text{-}score)^2 \times \sigma \times (1-\sigma)}{e}$$

with a *z-score* computed with 95% confidence, $e = 5\%$ margin of error, and standard deviation σ, which yielded 385 sample sets on which to train and test our models. We illustrate with 18 testing years because it is the maximum value of $\binom{36}{X}, X \in [1, 36]$.

Decision trees, and thus random forests, are scale-invariant [32]. This means that although our data varies greatly in scale between, for example, sea ice extent, in millions of km^2, and wind speeds, less than 1 m/s, the models' accuracy is unaffected. This is an advantage over many other ML models, and we can leave the data generally untouched. However, random forests extrapolate poorly for data outside of their training's minimum and maximum values [33]. This presented a problem for our analysis of the dataset because, as shown in Figure 1, the latter third of the data has values generally lower than any in the first two thirds. We detrended training and testing data separately to mitigate that problem by forcing the data to have a zero mean. After training and fitting our models, we retrended the data and the model's predictions to evaluate their error.

3.4. Model Training and Hyper-Parameter Tuning

Finally, we trained RFR models on the data the training splits provided. Note that the trees in our forests were allowed to grow until all leaves were pure, even if they contained a single sample. Decision trees are often pruned to reduce overfitting, but Breiman [34] suggests letting trees grow fully in random forests to boost accuracy and increase ensemble diversity. Banfield et al. [35, 27] also discuss ensemble size in random forests and conclude that many more trees are necessary than are typically used. Ensemble size is an important hyper-parameter to tune because the number of trees in the forest directly impacts the possible feature sets the forest can explore, and too many trees can reduce a random forest's performance while also sacrificing run-time. Our forests comprised 250 decision trees. The number of trees was determined empirically. Forests of size 10, 50, 100, 250, 500, and 1000 trees were evaluated and their performance was measured on the basis of the test $\overline{R^2}$ (average R^2) and average test anomaly correlation coefficient (ACC),

which are detailed in Section 3.6. We found that 250 tree models maximized $\overline{R^2}$ and $\overline{ACC}$. Lastly, the trees in each forest used mean squared error as their nodes' splitting criterion.

3.5. Feature Importance Measurement

We used Gini importance because of the non-linearities in climate data; in particular, Gini importance is not susceptible to data multicollinearities. Given that all of our features come from the same complex system, it would be difficult to eliminate features by simple correlation measures. In standard usage, Gini importance is normalized to compare relative importance within a single dataset. We chose to preserve the absolute importance values, letting us compare across datasets.

We also considered drop-column and permutation importance methods [14]. However, we found them to be unsuitable because they are highly susceptible to multicollinearity. Because many physical processes are directly acting on each other, Arctic features are inherently correlated, and any leave-one-out importance method will highlight that correlation. We found that the correlation leads these methods to attribute more importance to the least correlated feature, and it becomes difficult to glean meaningful insights.

3.6. Model Evaluation

We used the R^2 (coefficient of determination) from the Nash-Sutcliffe efficiency definition, given by:

$$R^2(\hat{y}, y) = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \overline{y})^2},$$

where y are the true values, $\hat{y}$ are the predicted values, and $\overline{y}$ is the mean of y. This definition has a range of $(-\inf, 1]$ where 1 is the best possible score.

In addition to $\overline{R^2}$, we evaluated model performance with average MAE ($\overline{MAE}$) and $\overline{ACC}$. Again, average here means the mean value measured in 385 models with random training and testing year splits. Since $\overline{MAE}$ is in millions of km^2, we took the Sea Ice Outlook's 2019 season report [36] as a baseline. This report includes several different types of data-driven models and presents one-year forecasts. These should have less error than ours, given how many more years we forecasted at once. With the exception of a few outliers between 2012 and 2019, sea ice forecast error was between -0.4 and 0.6 million km^2.

Figure 2: June feature importance. Standard box-and-whisker plot [37] of values for 13 predictions generated by 385 models. The average R^2, anomaly correlation coefficient (ACC), and mean absolute error (MAE) are displayed in the gray boxes. The blue line in each dataset is the mean importance of a random variable in each feature set.

ACC is the Pearson's correlation coefficient (r-value) of sea ice extent anomalies. A time series' anomaly is a measure of the data's deviation from its *climatology*. In our case, the climatology is the mean value of the true values the models are attempting to forecast. This function is defined by:

$$ACC(\hat{y}, y) = \frac{\sum[(\hat{y} - \overline{y})(y - \overline{y})]}{M \times \sigma_{\hat{y}} \times \sigma_y}$$

where y are the true values, $\hat{y}$ are the predicted values, M is the number of samples in y and $\hat{y}$, $\overline{y}$ is the mean or climatology of y, $\sigma_{\hat{y}}$ is the standard deviation of the predicted values, and σ_y is the standard deviation of the true values.

4. Results

Our goal is to learn the importance of climate features on the predictions made by E3SM and compare that to the actual importance of those features

Figure 3: July feature importance. Standard box-and-whisker plot [37] of values for 13 predictions generated by 385 models. The average R^2, anomaly correlation coefficient (ACC), and mean absolute error (MAE) are displayed in the gray boxes. The blue line in each dataset is the mean importance of a random variable in each feature set.

on observed sea ice extent. We found that was best accomplished by training RFRs on 23 uniformly randomly chosen years and testing with the remaining 13. Our performance measure was based on the mean of $\overline{R^2}$ scores among datasets for the June input data. This train-test-split resulted in maximum and minimum $\overline{R^2}$ scores of 0.88 and 0.77, respectively, yielding a measure of 0.83. $\overline{R^2}$ denotes the average R^2 of the 385 models.

We replicated our analysis for each month between June and August, predicting September SIE. Each subsequent month generates less error. Within each dataset, each feature's relative importance changes. Some features' importance is correlated with the progression of months, while others appear to change randomly.

Figure 2 shows June's feature importance values. The average train and test error values indicate that the models generally learn the data well. The blue line shows the mean feature importance of a random variable included in each model's feature set. The random variable indicates a lower bound on importance; any feature with an importance value near this line has virtually

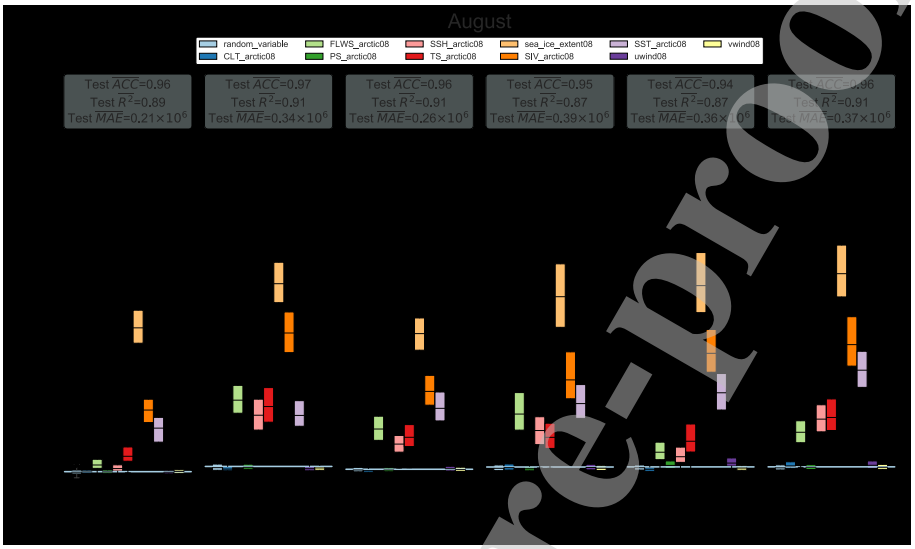

Figure 4: August feature importance. Standard box-and-whisker plot [37] of values for 13 predictions generated by 385 models. The average R^2, anomaly correlation coefficient (ACC), and mean absolute error (MAE) are displayed in the gray boxes. The blue line in each dataset is the mean importance of a random variable in each feature set.

no importance. We found that adding a random variable decreases individual model performance, but the effect is minimized when taking the mean over every model.

There are some similarities between each dataset. They share the same list of six important features, though their order and magnitudes differ. SIV is consistently the most important, though the degree of absolute importance varies. SIV, TS, SSH, SIE, FLWS, and SST are important in each dataset. The datasets, except for simulation 3, share the same list of unimportant features as well. These are CLT, PS, uwind, and vwind. One apparent exception is June's PS in Figure 2: simulation 3; however, excluding PS from the training data, results in a negligible difference in $\overline{R^2}$ (0.7681 vs. 0.7682).

July features, shown in Figure 3, predicted as well or better than June in each of our error metrics; simulation 3 had the lowest $\overline{R^2}$, 0.78, and simulation 2 had the highest, 0.88. The same features were important in July as in June, but the relative importance values changed. June's sea ice extent became

more important in the observed dataset, surpassing the importance of SIV. SSH became less important in the observed dataset, too, settling just above the random variable. SSH remained as important in the simulation datasets.

The most dramatic change in importance occurs in August. These results are in Figure 4. Error was significantly better with simulations 3 and 4 having the minimum $\overline{R^2}$, 0.87, and simulations 1 and 2 having the maximum, 0.91. In August, sea ice extent was always the most important. The importance values of the remaining features generally changed very little throughout datasets.

5. Discussion

We found that our RFR ML models were able to accurately learn each of the datasets. After examining the Gini importances computed within each model, we discovered some key differences in how each dataset relates to September pan-Arctic sea ice extent.

A problem with our dataset is that the satellite record only goes back to 1979. One solution is to adapt the models to forecast sea ice extent continuously throughout each year. This is in line with Reid and Tarantino's approach [18] (see Section 2), but with random forests instead of support vector machines and including many features instead of only sea ice extent. The models would train on the full year of data and see 432 data points rather than 36 in the time series. Several observed features are measured more frequently than monthly, some every few hours of every day, so a means to incorporate inconsistent sampling resolutions of features should be investigated to leverage all of the data available. Another solution could be to use a surrogate model to generate more data that is similar to the first 15 years of observed data, which have a much flatter trend. The surrogate model would let the new data agree with what the model learns about input features.

The combined error metrics and general consistency of results between each dataset suggests that our models have learned the data well, and the feature analysis can identify key patterns. It is meaningful that the same six features are considered important across datasets and input-months. Since our analysis is of the pan-Arctic region, it is possible that the set of unimportant features would be more important in specific subregions of the Arctic.

Though the most important feature in June and August is consistent between simulation and observation, the absolute importance differs markedly. One clear pattern is that June shows an acute reliance on sea ice volume for

both observations and simulations. By August the reliance is traded for sea ice extent. This finding is consistent with earlier studies evaluating sea ice predictability using lag-correlation analyses with ESM ensemble data [38, 39].

Although the observed and simulated data share patterns, there is a clear difference between them. In July, simulations and observed data do not agree on the most important feature. In June, July, and August, simulated data relies too heavily on almost all the important features. In each dataset, importance values diminish for the remaining features in June and July, and their distributions overlap more than they did in June, but the observed dataset still shows the least importance in FLWS, SSH, TS, and SST.

Interestingly, simulations 1 and 2 forecasted with the highest $\overline{R^2}$ each input month, and simulations 3 and 4 had the lowest $\overline{R^2}$ in each input month. Simulations 1 and 2 have the lowest $\overline{MAE}$ and highest $\overline{ACC}$ among the simulation runs, and 3 and 4 have the highest $\overline{MAE}$ and lowest $\overline{ACC}$ among the simulations runs. Although the differences are small, these consistencies may indicate some commonality between these simulation runs.

Our ML models performed better on the observed data than on the simulations as measured by $\overline{MAE}$ and $\overline{ACC}$, but is not reflected in $\overline{R^2}$. That suggests that the mean value, or the trend after retrending, was very predictable, but its intervariability, which R^2 explains, was less predictable. The likely explanation is in the difference in the complexity of the systems. Observed features of the continuous Earth system are artificially discretized. In any complex system, intervariability is difficult to forecast. However, because we chose largely relevant features as predictors, we could capture the macro-level patterns, as evidenced by the macro-level error measures: $\overline{MAE}$ and $\overline{ACC}$.

6. Conclusions

We demonstrated that random forest regression and the associated Gini importance measure can provide insight into why ESMs incorrectly estimate sea ice extent in recent decades. We found a discrepancy in the feature importance between observed and simulation datasets. In particular, the discrepancy between E3SM and observation appear to be due to an over-reliance on June sea ice extent and August sea ice volume. The order of feature importance was also different between E3SM and observation, and the ordering was not consistent within E3SM ensemble members. In all cases, E3SM over-relies on six features compared to observed data. Machine

learning allows us to fill the gaps in the underlying physics of ESMs, providing a metric for Stroeve et al.'s [16] hypothesis that ESMs are missing complex relations and causal mechanisms.

In the future, we can evaluate more features that can be measured or constructed in each dataset. An analysis, including all months of the year in each model will be elucidating as well. Sea ice extent is measured daily via satellite imagery. We can understand how each dataset explains sea ice extent at a higher resolution every month of the year.

We can repeat our analysis on other regions, including Antarctica, where there are also problematic disagreements with observations [8]. An analysis like this of other climate models could be insightful too. It would be particularly interesting to compare simulations in which there few to no correlated features. That would allow for variations on the analysis, such as more modeling approaches, which require linearly independent features, and more feature analysis methods, such as drop-column importance, which would otherwise struggle with multicollinearities.

Further insight could be gained by repeating our analysis with a machine learning method other than RFR, however the following methods have their own challenges. Most neural network models would need more observed data than is available to converge. We found that multiple linear regression cannot learn the data well because the relationships between features are nonlinear. Reid and Tarantino [18] found that SVR can forecast the data well, but it is unclear what the best feature analysis method would be.

Given the discoveries in this paper, we can run experiments with E3SM to determine how reducing feature disagreements between the observed and simulation datasets impact E3SM's forecasts. That process may not yield results for several reasons, including that E3SM's real feature set is large and complex, focusing analysis on the Arctic region is too restricting to estimate the effects of the global Earth model, or our ML models are too limited by small datasets. Despite these challenges, our results can potentially guide climate modelers as they develop the next generation of ESMs.

Acknowledgments

This work is supported by Sandia Earth Science Investment Area Laboratory Directed Research and Development funding. Sandia National Laboratories is a multimission laboratory managed and operated by National

Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525.

References

[1] J. Stroeve, D. Notz, Changing state of Arctic sea ice across all seasons (sep 2018). doi:10.1088/1748-9326/aade56.

[2] Arctic report card 2019, Tech. rep. (2019).
URL https://www.arctic.noaa.gov/Report-Card

[3] L. C. Smith, S. R. Stephenson, New trans-Arctic shipping routes navigable by midcentury, PNAS 110 (13) (2013) 4871–4872.

[4] H. Goosse, J. E. Kay, K. C. Armour, A. Bodas-Salcedo, H. Chepfer, D. Docquier, et al., Quantifying climate feedbacks in polar regions, Nature Communications 9 (1919) (2018).

[5] F. Sevellec, A. V. Fedorov, W. Liu, Arctic sea-ice decline weakens the atlantic meridional overturning circulation, Nature Climate Change 7 (2017) 604–610.

[6] J. Cohen, K. Pfeiffer, J. A. Francis, Warm Arctic episodes linked with increased frequency of extreme winter weather in the United States, Nature Communications 9 (869) (2018).

[7] I. Cvijanovic, B. D. Santer, C. Bonfils, D. D. Lucas, J. C. H. Chiang, S. Zimmerman, Future loss of Arctic sea-ice cover could drive a substantial decrease in california's rainfall, Nature Communications 8 (1947) (2017).

[8] E. Rosenblum, I. Eisenman, Sea ice trends in climate models only accurate in runs with biased global warming, Journal of Climate 30 (16) (2017) 6265–6278. doi:10.1175/JCLI-D-16-0455.1.

[9] A. G. Meehl, C. Covey, T. Delworth, M. Latif, B. Mcavaney, J. F. B. Mitchell, et al., THE WCRP CMIP3 Multimodel Dataset: A New Era in Climate Change Research, American Meteorological Society (September) (2007). doi:https://doi.org/10.1175/BAMS-88-9-1383.

[10] J. Stroeve, M. M. Holland, W. Meier, T. Scambos, M. Serreze, Arctic sea ice decline: Faster than forecast, Geophysical Research Letters 34 (9) (2007). doi:10.1029/2007GL029703.

[11] M. G. A. Taylor Karl E., Stouffer Ronald J., An Overview of CMIP5 and the Experiment Design, American Meteorological Society 3 (april) (2012) 485–498. doi:10.1175/BAMS-D-11-00094.1.

[12] E3SM Project, Energy Exascale Earth System Model (E3SM), [Computer Software] https://dx.doi.org/10.11578/E3SM/dc.20180418.36 (Apr. 2018). doi:10.11578/E3SM/dc.20180418.36.
URL https://dx.doi.org/10.11578/E3SM/dc.20180418.36

[13] V. Eyring, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, et al., Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geoscientific Model Development 9 (5) (2016) 1937–1958. doi:10.5194/gmd-9-1937-2016.
URL https://www.geosci-model-dev.net/9/1937/2016/

[14] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32. doi:10.1023/A:1010933404324.
URL http://dx.doi.org/10.1023/A%3A1010933404324

[15] S. Nembrini, I. R. Ko, M. N. Wright, C. Lu, The revival of the Gini importance? 34 (May) (2018) 3711–3718. doi:10.1093/bioinformatics/bty373.

[16] J. C. Stroeve, V. Kattsov, A. Barrett, M. Serreze, T. Pavlova, M. Holland, et al., Trends in Arctic sea ice extent from CMIP5, CMIP3 and observations, Geophysical Research Letters 39 (16) (2012) 1–7. doi:10.1029/2012GL052676.

[17] M. Ionita, K. Grosfeld, P. Scholz, R. Treffeisen, G. Lohmann, September Arctic Sea Ice minimum prediction - a new skillful statistical approach, Earth System Dynamics Discussions (2018) 1–23doi:10.5194/esd-2018-61.
URL https://www.earth-syst-dynam-discuss.net/esd-2018-61/

[18] T. G. Reid, P. M. Tarantino, Arctic sea ice extent forecasting using support vector regression, in: Proceedings - 2014 13th International Conference on Machine Learning and Applications, ICMLA 2014,

Institute of Electrical and Electronics Engineers Inc., 2014, pp. 1–6. doi:10.1109/ICMLA.2014.7.

[19] G. Peng, W. N. Meier, D. J. Scott, M. H. Savoie, N. Snow, A long-term and reproducible passive microwave sea ice concentration data record for climate studies and monitoring (2013) 311–318doi:10.5194/essd-5-311-2013.

[20] A. Schweiger, R. Lindsay, J. Zhang, M. Steele, H. Stern, R. Kwok, Uncertainty in modeled Arctic sea ice volume, Journal of Geophysical Research: Oceans 116 (9) (2011) 1–21. doi:10.1029/2011JC007084.

[21] NOAA, OAR, ESRL-PSD, Ncep-doe reanalysis 2, nCEP_Reanalysis 2 data provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA (2019).
URL https://www.esrl.noaa.gov/psd/

[22] NOAA, OAR, ESRL-PSD, Noaa extended reconstructed sea surface temperature, NOAA_ERSST_V4 data provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA (2019).
URL https://www.esrl.noaa.gov/psd/

[23] J.-C. Golaz, P. M. Caldwell, L. P. V. Roekel, M. R. Petersen, Q. Tang, J. D. Wolfe, et al., The DOE E3SM Coupled Model Version 1 : Overview and Evaluation at Standard Resolution (2019). doi:10.1029/2018MS001603.

[24] J. E. Kay, C. Deser, A. Phillips, A. Mai, C. Hannay, G. Strand, et al., The community earth system model (cesm) large ensemble project: A community resource for studying climate change in the presence of internal climate variability, Bulletin of the American Meteorological Society 96 (8) (2015) 1333–1349. arXiv:https://doi.org/10.1175/BAMS-D-13-00255.1, doi:10.1175/BAMS-D-13-00255.1.
URL https://doi.org/10.1175/BAMS-D-13-00255.1

[25] B. J. Zib, X. Dong, B. Xi, A. Kennedy, Evaluation and intercomparison of cloud fraction and radiative fluxes in recent reanalyses over the arctic using BSRN surface observations, Journal of Climate 25 (7) (2012) 2291–2305. doi:10.1175/JCLI-D-11-00147.1.

[26] L. Breiman, Bagging predictors, Machine learning 24 (2) (1996) 123–140.

[27] R. E. Banfield, L. O. Hall, K. W. Bowyer, W. P. Kegelmeyer, A Comparison of Decision Tree Ensemble Creation Techniques 29 (1) (2007) 173–180.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[29] L. Breiman, Arcing classifiers, Ann. Statist. 26 (3) (1998) 801–849. doi:10.1214/aos/1024691079.
URL https://doi.org/10.1214/aos/1024691079

[30] S. Van Der Walt, S. C. Colbert, G. Varoquaux, The numpy array: a structure for efficient numerical computation, Computing in Science & Engineering 13 (2) (2011) 22.

[31] M. Waskom, the seaborn development team, mwaskom/seaborn (Sep. 2020). doi:10.5281/zenodo.592845.
URL https://doi.org/10.5281/zenodo.592845

[32] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, Classification and Regression Trees, Wadsworth International Group, Belmont, California, 1984.

[33] T. Hengl, M. Nussbaum, M. N. Wright, G. Heuvelink, B. Gräler, Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables, PeerJ 6 (2018). doi:10.7717/peerj.5518.
URL https://doi.org/10.7717/peerj.5518

[34] L. Breiman, Rejoinder: Arcing classifiers, The Annals of Statistics 26 (3) (1998) 841–849.
URL http://www.jstor.org/stable/120059

[35] R. E. Banfield, L. O. Hall, K. W. Bowyer, W. P. Kegelmeyer, A new ensemble diversity measure applied to thinning ensembles, in: T. Windeatt, F. Roli (Eds.), Multiple Classifier Systems, Springer Berlin Heidelberg, Berlin, Heidelberg, 2003, pp. 306–316.

[36] U. S. Bhatt, P. Bieniek, C. Bitz, E. Blanchard-Wrigglesworth, H. Eicken, H. Goessling, et al., 2019 sea ice outlook full post-season report, editors: Turner-Bogren, B. and H. V. Wiggins (February 2020).
URL https://www.arcus.org/sipn/sea-ice-outlook/2019/post-season

[37] R. McGill, J. W. Tukey, W. A. Larsen, Variations of box plots, The American Statistician 32 (1) (1978) 12–16.

[38] A. C. Ordonez, C. M. Bitz, E. Blanchard-Wrigglesworth, Processes controlling Arctic and Antarctic sea ice predictability in the Community Earth System Model, Journal of Climate 31 (2018) 9771–9786.

[39] E. Blanchard-Wrigglesworth, K. C. Armour, C. M. Bitz, E. Deweaver, Persistence and inherent predictability of arctic sea ice in a GCM ensemble and observations, Journal of Climate 24 (1) (2011) 231–250. doi:10.1175/2010JCLI3775.1.