



<https://doi.org/10.1038/s43247-021-00225-4>

OPEN

## Training machine learning models on climate model output yields skillful interpretable seasonal precipitation forecasts

Peter B. Gibson<sup>1</sup>, William E. Chapman<sup>1</sup>, Alphan Altinok<sup>2</sup>, Luca Delle Monache<sup>1</sup>, Michael J. DeFlorio<sup>1</sup> & Duane E. Waliser<sup>2</sup>

A barrier to utilizing machine learning in seasonal forecasting applications is the limited sample size of observational data for model training. To circumvent this issue, here we explore the feasibility of training various machine learning approaches on a large climate model ensemble, providing a long training set with physically consistent model realizations. After training on thousands of seasons of climate model simulations, the machine learning models are tested for producing seasonal forecasts across the historical observational period (1980-2020). For forecasting large-scale spatial patterns of precipitation across the western United States, here we show that these machine learning-based models are capable of competing with or outperforming existing dynamical models from the North American Multi Model Ensemble. We further show that this approach need not be considered a ‘black box’ by utilizing machine learning interpretability methods to identify the relevant physical processes that lead to prediction skill.

<sup>1</sup>Center for Western Weather and Water Extremes (CW3E), Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA, USA.

<sup>2</sup>NASA Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA. ✉email: [peter.gibson@niwa.co.nz](mailto:peter.gibson@niwa.co.nz)

**Sources of seasonal predictability.** The climatology and variability of precipitation across the western United States present a unique seasonal forecasting challenge. Relatively low precipitation totals combined with high year-to-year variability are often received in the form of a relatively small number of atmospheric rivers across winter months<sup>1</sup>. Individually, these storms have proven challenging to forecast at lead times beyond the weather time horizon<sup>2,3</sup>. During the recent severe California drought (years 2012–2016), the challenges for decision-makers under forecast uncertainty were highlighted. As widely documented, the expected positive anomaly of precipitation across California and the Southwest under the major El Niño event of 2015/2016 did not eventuate as anticipated, and instead the devastating drought continued<sup>4</sup>. Given that the economic costs of severe drought can frequently exceed \$1B annually across California<sup>5,6</sup>, improving the skill of seasonal precipitation forecasts remains a top priority for water resource managers.

In a seasonal forecasting context, teleconnections are best viewed as probabilistically loading the dice in favor of a certain outcome (i.e., dry versus wet conditions). The El Niño Southern Oscillation (ENSO) is known to be the primary driver of seasonal forecast skill across North America<sup>7–10</sup>, yet its signal-to-noise ratio is such that unexpected outcomes will occasionally occur by chance<sup>4,11,12</sup>. Other studies have shown that traditional indices for describing ENSO variability (i.e., Niño3.4) may not be optimal for capturing the teleconnection to western US precipitation<sup>13</sup>. Studies using both model simulations and observations have also shown that tropical diabatic heating anomalies in key regions across the western tropical Pacific, at times independent from ENSO, substantially increase the likelihood of ridging and subsequently drought conditions across California<sup>14,15</sup>.

The Indian Ocean has been suggested to play a role in mediating the ENSO precipitation teleconnection to North America in certain years<sup>16</sup>. While the tropospheric mid-latitude jet stream has a much shorter memory than tropical sea-surface temperatures (SSTs), certain configurations of the jet have also been shown to be predictable across subseasonal timescales, which has particular relevance for regional precipitation predictability over the western US<sup>11</sup>. In the stratosphere, both tropical and polar stratospheric variability can have appreciable impacts on precipitation across North America, offering a potential source of predictability on subseasonal-to-seasonal timescales<sup>17,18</sup>. To take full advantage of these potential sources of seasonal predictability, forecast models must capture a wide range of these teleconnections and their possible interactions.

**Existing approaches to seasonal forecasting.** Seasonal forecasting methods can be broadly categorized into dynamical, empirical (i.e., statistical or machine learning), or hybrid-based (i.e., dynamical models combined with empirical approaches). Seasonal forecasts using dynamical models typically involve probabilistic forecasts derived from the spread of the ensemble members, which relates to very small uncertainties in the model initial conditions that grow rapidly with time. The ensemble mean of these individual ensemble members can be used to forecast any signal that might emerge beyond the noise of individual weather events. The North American Model Ensemble Project (NMME)<sup>19,20</sup> has provided an opportunity to estimate the skill of start-of-the-art dynamical reforecasts. Seasonal forecast skill for precipitation across the western US has generally been found to be low after 2-weeks lead-time, but with dry extremes better forecasted than wet extremes<sup>21</sup>. Recently, seasonal forecast skill has been evaluated across different suites of NMME models that represent upgrades across model versions<sup>20</sup>. While seasonal

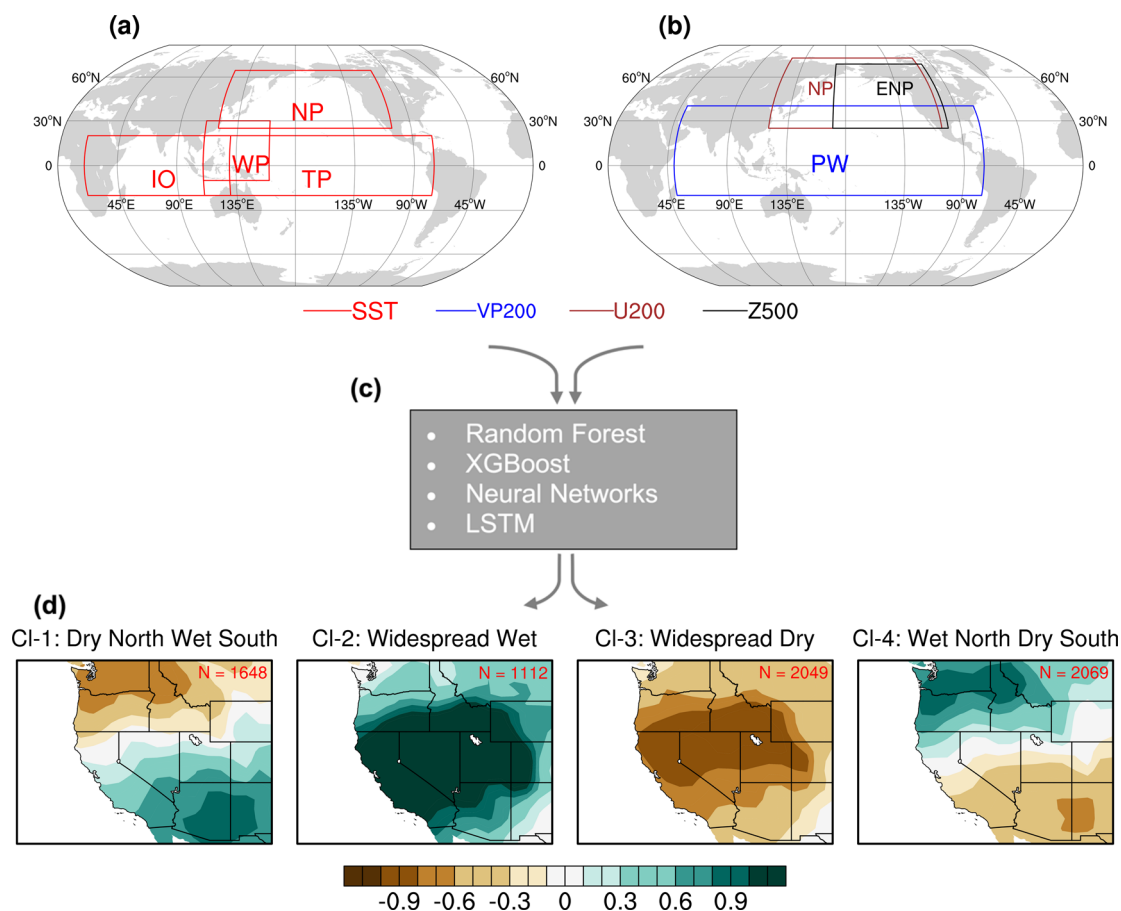
temperature forecast skill was shown to improve across successive upgrades, minimal improvements were found for precipitation suggesting a potential saturation of skill for models run at this resolution.

Statistical seasonal forecasts of precipitation in the United States have a long history, and often implement classical canonical correlation analysis (CCA)<sup>7,22</sup>. This approach models linear relationships between two sets of predictor variables, commonly between lagged spatial fields of SST and temperature or precipitation. Within the classical CCA framework, it is difficult to include multiple predictor variables and their interactions without overfitting<sup>23</sup>, the temporal nature of the data is not explicitly modeled beyond the use of lagged correlations, and only linear relationships through correlation are directly captured. Despite the large ongoing investments in developing dynamical model ensembles, a relatively small amount of research has explored recent advances in machine learning for improving seasonal forecast skill<sup>24</sup>.

Another important barrier to implementing empirical approaches for seasonal forecasting is the limited sample size of observational data needed in model training<sup>25</sup>, which impacts both traditional statistical approaches and machine learning-based approaches. For example, for reliable modeling of non-linear interactions between multiple predictor variables, an extensively large number of cases per predictor variable is required to limit overfitting<sup>26</sup>. Since this is clearly not possible with the limited record afforded by current observational or reanalysis products (~40–100 years), one promising alternative involves a hybrid-based approach of training machine learning models on large climate models ensembles<sup>25,27,28</sup>. This approach can greatly increase the training dataset sample size to span several thousand seasons and has achieved skillful ENSO predictions at lead times exceeding 1-year through training convolutional neural networks (CNN) on historical climate model simulations<sup>25</sup>. Other studies have also achieved skillful seasonal forecasts through applying relatively simple regularized regression models to climate model simulations<sup>27,28</sup>. In this study, through testing a wider range of machine learning models, we build upon these prior applications of training statistical and machine learning models on long climate model simulations for the purpose of seasonal prediction (Fig. 1). We implement cluster analysis to target the more predictable large-scale spatial patterns of precipitation, contributing to the observed seasonal prediction skill. Lastly, we implement a range of interpretable machine learning approaches to identify the relevant physical processes that contribute to prediction skill.

## Results and discussion

**Classification accuracy.** After training and calibrating each machine learning model on Community Earth System Model Large Ensemble (CESM-LENS) data (see Section “Climate model training data”, and Fig. 1) the models’ configurations were frozen and used to forecast seasonal precipitation clusters across the observational record. To do so, the models are driven by input data based on the observed atmospheric and oceanic conditions prior to the target season (e.g., using October and earlier conditions to predict November through January (NDJ)). This ‘test set’ evaluation of the models provides an estimate of future model performance since the models are making predictions on data unseen in the training phase. The classification accuracy of four machine learning models tested is shown in Fig. 2a for NDJ seasonal predictions and in Fig. 2b for January through March (JFM) seasonal predictions. The machine learning accuracy (red bars) is presented alongside the NMME accuracy (white bars) and ensemble accuracy (blue bars), presented for the overlapping test

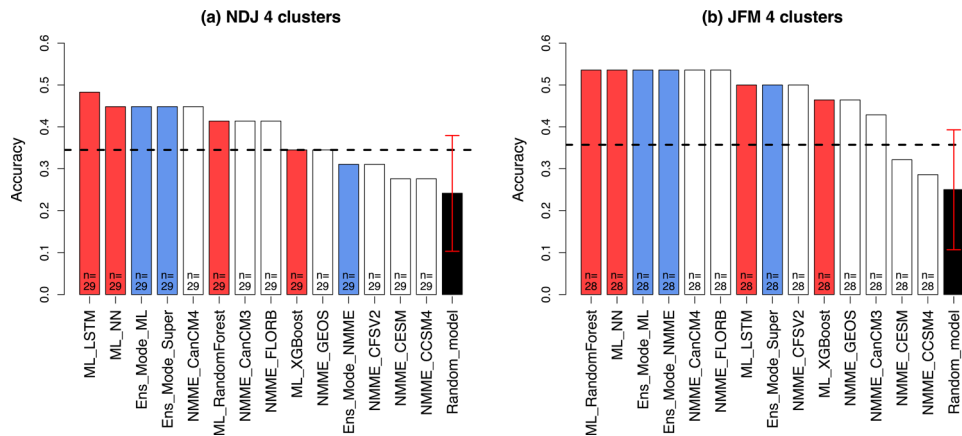


**Fig. 1 Summary of methodology for seasonal cluster prediction.** Panels **a** and **b** show lagged predictor variable regions from which EOFs were derived. Panel **a** shows ocean-based (SST) regions, panel **b** shows atmosphere-based (VP200, U200, and Z500) regions, with variable names referred to in the text. Panel **c** shows the four machine learning methods trained on CESM-LENS data/regions from panels (**a** and **b**). Panel **d** shows the predictand clusters derived from K-means clustering of 3-month standardized precipitation anomalies over the western US. The sample size (number of seasons) is shown for each cluster from CESM-LENS data. Models are trained separately for making seasonal predictions of November through January (NDJ) and January through March (JFM) clusters, using previous months as predictors.

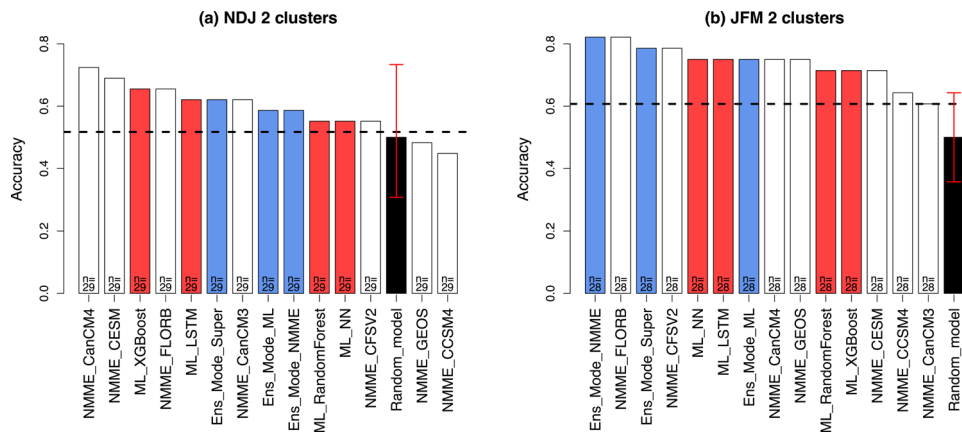
period years. For the ensemble methods, *Ens\_Mode\_ML* is calculated as the ensemble mode prediction from the four machine learning models, *Ens\_Mode\_NMME* is calculated as the ensemble mode prediction from the seven NMME models, and *Ens\_Mode\_Super* is calculated as the ensemble mode prediction from all NMME and machine learning models.

For the NDJ predictions, three of the machine learning models (LSTM, NN, RF) have accuracy in the 40–50% range. While this accuracy remains somewhat modest, it is skillful relative to both baseline methods tested: a random guess model and the most frequent cluster prediction. Furthermore, for NDJ, two of these machine learning models have accuracy above or equal to each of the NMME models tested, and the ensemble prediction accuracy from the machine learning models (*Ens\_Mode\_ML*) exceeds that from the NMME models (*Ens\_Mode\_NMME*). The classification accuracy for JFM, across the latter half of the water year, is generally found to be improved in the majority of models compared to NDJ, consistent with the previous studies<sup>29</sup>. The Random Forest and NN models are found to be top performers in JFM, with classification accuracy >50%. A number of NMME models also show skill relative to baseline methods in JFM, most notably CanCM4 and FLOBR which are also competitive with the machine learning models.

In examining misclassifications in various models, we found a general tendency to misclassify the cluster based on a failure to predict the precise positioning of the anomaly dipole. For example, a widespread dry pattern (cluster 3, Fig. 1) may have a greater tendency to be misclassified into the wet north dry south pattern (cluster 4). Depending on the application, this forecast error is likely less critical as the sign of the anomaly is still correctly forecasted across a large proportion of the Southwest under dry conditions. On the other hand, forecasting widespread dry (cluster 3) under a widespread wet occurrence (cluster 2) can clearly be considered a more problematic forecast. To explore this further, in model post-processing, we computed the accuracy in Fig. 3 after grouping cluster 1 and cluster 2 together (wet southwest group), and cluster 3 and 4 together (dry southwest group). As expected, this additional cluster grouping results in accuracy improvements in both individual models as well as the baseline methods. In JFM, a number of individual models (both machine learning and NMME) and their ensembles display accuracy in the 70–80% range which is skillful relative to both baselines. For NDJ, individual model accuracy can approach 60–70%, but in the case of NDJ, this level of accuracy is not skillful relative to the random guess baseline. Notably, for JFM, this highlights the increase in classification accuracy that can be



**Fig. 2 Accuracy of machine learning models and NMME models.** Accuracy of machine learning models (red), NMME models (white), and ensemble models (blue) for NDJ (panel **a**) and JFM (panel **b**) seasons. Accuracy is defined as the proportion of correct predictions. The sample size (number of predictions made) is given at the base of each bar. Baseline skill is defined here in two ways: (1) the horizontal line defined by the frequency of the most common cluster; (2) a random model prediction repeated 1000 times with bars showing the 5th/95th percentile of the random prediction accuracy. The Ensemble models (blue) are based on the ensemble mode cluster prediction across their respective groups, with Ens\_mode\_Super based on the ensemble mode across all models. Here accuracy is computed across overlapping years between all models in the test set, refer to Supplementary Material Fig. S11 for accuracy across all years.



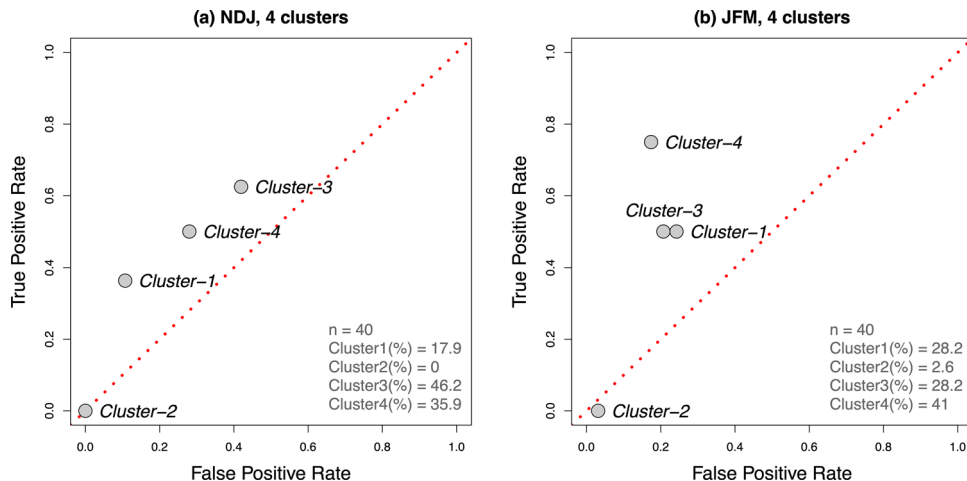
**Fig. 3 Improved accuracy from combining clusters.** As in Fig. 2 but here accuracy is computed after clusters have been grouped together with clusters 1,2 forming a new cluster (broadly representing a wet southwest), and clusters 3,4 forming a new cluster (broadly representing a dry southwest) for NDJ (panel **a**) and JFM (panel **b**) seasons. Refer to Fig. 1 for cluster patterns. Here accuracy is computed across overlapping years between models, refer to Supplementary Material Fig. S12 for accuracy across all years.

achieved by reducing the spatial precision in the target prediction variable. Similar results have been documented in other studies, where aggregating over larger regions has generally been shown to significantly increase the skill by relaxing the predictand spatial requirement and avoiding direct predictions across the smaller less predictable components<sup>30–32</sup>.

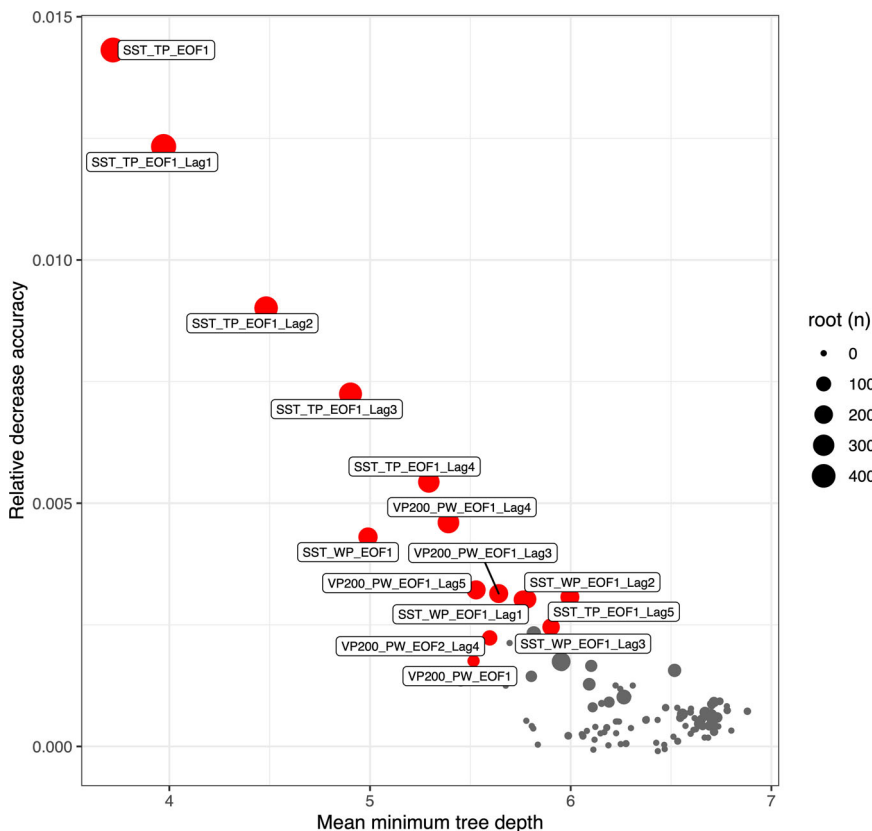
To investigate whether certain clusters are typically better predicted than others, Fig. 4 presents a receiver operating characteristic (ROC) diagram for NDJ and JFM seasons separated by cluster. The analysis is presented for predictions from the RF model, shown to be one of the top-performing models across JFM (Fig. 2b, Supplementary Material Fig. S11). More predictable clusters in the ROC diagram are indicated by larger true positive rates and smaller false-positive rates. In both seasons, cluster 4 (wet north dry south pattern) and cluster 3 (widespread dry) are generally found to be the most predictable patterns. In contrast, the model very rarely predicted the occurrence of the widespread wet cluster 2 (0% of predicted clusters in NDJ and 2.6% in JFM), despite the occurrence of cluster 2 in both the test set (15% in

NDJ and 17.5% in JFM, Supplementary Material Table S5) and in the training set (15.7% in NDJ and 16.4% in JFM). All other machine learning models also displayed a reduction in skill for predicting this cluster. Since cluster 2 was the least frequent cluster in the training dataset (Fig. 1d), we tested various approaches across the four machine learning models for increasing the predicted frequency of this cluster. However, modifying the class weights during training and stratified sampling approaches were not found to systematically increase the prediction skill for this cluster across models.

We suggest that the competitive machine learning accuracy results reported here largely stem from: (1) including a large pool of candidate predictor variables and accounting for non-linear relationships; (2) predicting smoother more predictable components of seasonal precipitation through the clusters obtained from K-means clustering. The use of clusters as the predictand has allowed investigation of skill on a per-cluster basis, providing insight into forecast errors in terms of the spatial pattern and sign of the precipitation field being forecast. Furthermore, we suggest



**Fig. 4 ROC accuracy by cluster.** ROC diagram for cluster predictions of NDJ season (panel a) and JFM season (panel b) from the Random Forest model on the test dataset (years 1980–2020). The percentage of predictions assigned to each cluster is given in the bottom right of each plot.

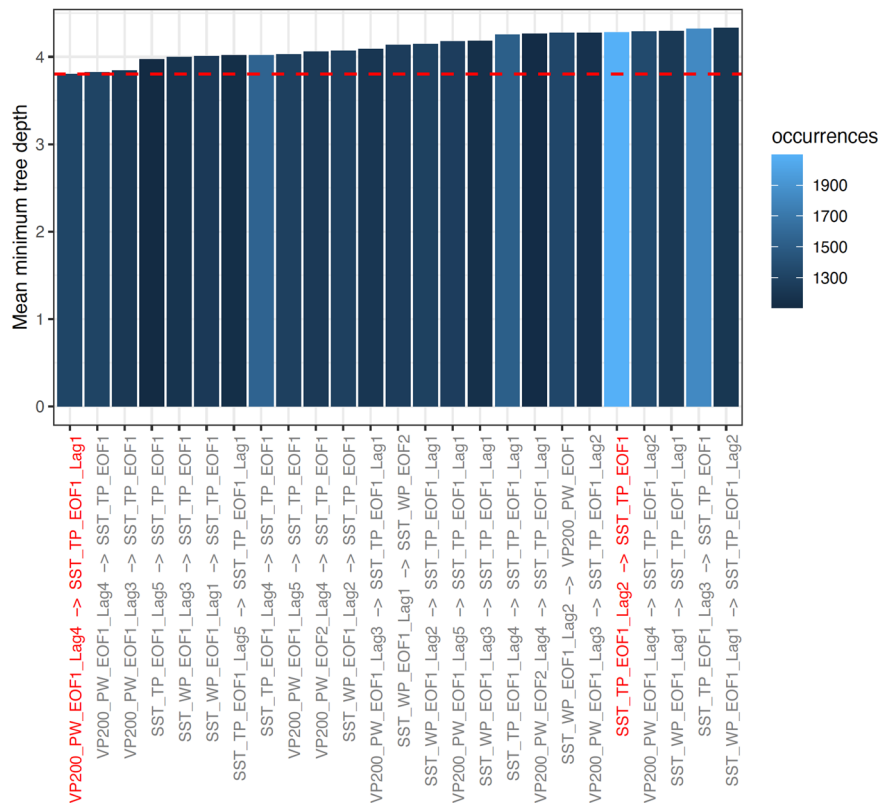


**Fig. 5 Individual variable importance for the Random Forest model.** Predictor variable importance plot for the Random Forest model predicting JFM clusters. Three factors were considered: (1) Relative decrease accuracy—the relative penalty to accuracy from shuffling each predictor variable; (2) mean minimum tree depth—the average minimum depth of the predictor variable across all decision trees in the Random Forest; (3) root—the number of instances when the predictor variable is positioned at the root of the decision tree, summed across all decision trees in the Random Forest. The 15 most important predictor variables are highlighted in red. Predictor variable labels relate to those in Fig. 1a, b. For example, SST\_TP\_EOF1 is the first EOF of tropical Pacific SST (i.e., December value) and SST\_TP\_EOF1\_Lag1 is the additional 1-month lag of the first EOF of tropical Pacific SST (November value).

that searching for skill in these larger-scale precipitation clusters, as opposed to attempting to predict seasonal precipitation on individual grid cells, is more closely aligned with the spatial scales of the dominant sources of predictability, namely the general positioning of ridges and troughs along stationary Rossby wave

trains. In the following sections, we explore the physical plausibility of the associations learned in the model training.

**Interpreting the Random Forest model.** An often-cited criticism of machine learning is the challenge of interpretability compared



**Fig. 6 Most important variable interactions for the Random Forest model.** Mean minimum tree depth for the top-25 pairwise variable interactions for the Random Forest model predicting JFM clusters. The variable interactions shown are the shallowest of all possible variable interactions, suggesting overall higher importance for influencing the probability of a prediction. Of these interactions, the shallowest and most frequent pairwise interactions are highlighted in red.

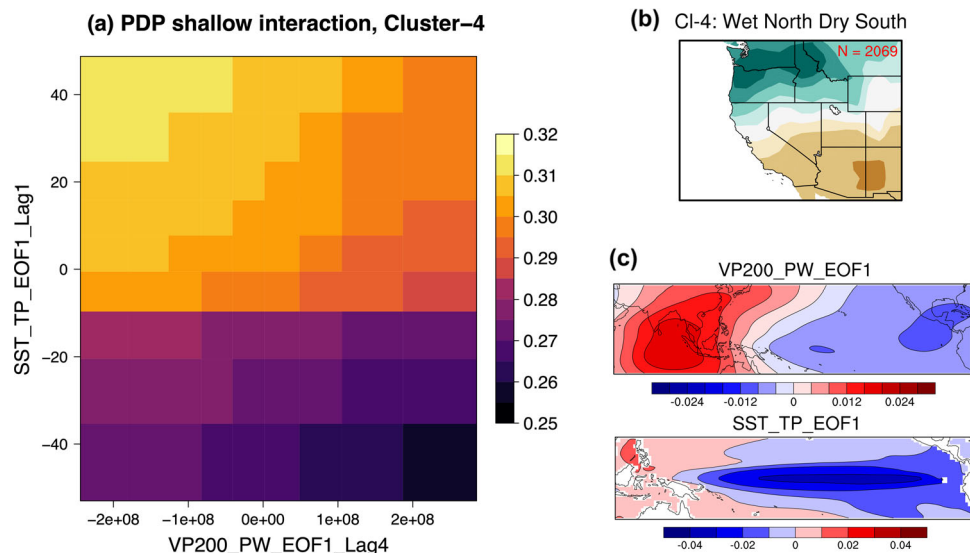
to much simpler linear models. The potential lack of interpretability has implications for the perceived credibility of the model, where machine learning models may achieve promising results for the wrong reasons<sup>33</sup>. In this section, we present further analysis that attempts to look inside the machine learning “black box”. We focus on the RF model, which was shown to be a top performer in predicting JFM clusters, but note that a number of these approaches are model-agnostic (e.g., permutation importance, Partial dependence plots, ALE plots, and LIME) and could be applied to the other machine learning models in future work.

The relative importance of individual predictor variables in the RF model is shown in Fig. 5, based on considering three variable importance measures. Overall, based on these measures, the most important predictor variables for JFM seasonal precipitation in the model are tropical Pacific SST anomalies from July through December (SST\_TP\_EOF1 Lag 0–5), velocity potential anomalies in the tropics from July through December (VP200\_PW\_EOF1 Lag 0-5 and VP200\_PW\_EOF2 Lag 4), and western tropical Pacific SST anomalies from September through December (SST\_WP\_EOF1 Lag 0–3). There is consistency between the three variable importance measures detailed in Fig. 5, providing further confidence that these predictor variables are indeed providing robust measures of importance. In particular, there is a general negative relationship between the *relative decrease accuracy* and *mean minimum tree depth* variables, indicating that the variables most important to classification accuracy are also generally positioned closer to the root of the decision tree, as expected. It is also noteworthy that 14 of the 15 top predictor variables correspond to EOF1. Since EOF1 by definition explains

the greatest variance, the fact that the Random Forest is capable of distinguishing these components as being more important is encouraging. Furthermore, the RF typically favors the lower lags of the predictor variables (i.e., lags 0–3) as opposed to lags 11 or 12. This aligns with the intuition that conditions closer to the target in time should carry more predictive information.

Notably, the top predictor variables highlighted in Fig. 5 are consistent with the current physical understanding of the dominant contributions to western US seasonal precipitation. A substantial body of work has highlighted the importance of ENSO as the primary driver of seasonal forecast skill over North America<sup>4,7,8</sup>. The fact that the first empirical orthogonal function (EOF) of tropical Pacific SST, that most closely related to ENSO variability, is found by the Random Forest model to be the most important predictor variable (Fig. 5) provides confidence in the ability of the model to distinguish the most relevant teleconnections from a large pool of candidate predictor variables. Western tropical Pacific SST variability has also been shown, in both modeling<sup>14</sup> and observational studies<sup>15</sup>, to be particularly important in driving a Rossby wave train response in the mid-latitudes and subsequently placing a ridge over the coastal western US that is typically associated with widespread drought. Western Pacific SST variables and their lags are prevalent among the top predictor variables in Fig. 5. Later in this Section, we further illustrate that the direction of the western tropical Pacific precipitation relationship is also consistent with past research, and how the association can be modulated by ENSO strength and variability.

Other studies have highlighted that the representation of ENSO by the Niño3.4 index is likely not optimal for capturing the



**Fig. 7** Partial dependence plot for the shallowest pairwise variable interaction. Panel **a** indicates that strongly negative values of VP200\_PW\_EOF1\_Lag4 (i.e., the August value of the first EOF of VP200) combined with a strongly positive SST\_TP\_EOF1\_Lag1 (i.e., the November value of the first EOF of tropical Pacific SST) increases the probability (shaded colors) of a JFM Cluster-4 (panel **b**) precipitation anomaly occurrence. This relates to La Niña like conditions in November and enhanced convection conditions in the Indian Ocean/Maritime continent in August (panel **c**). Partial dependence plots for other variable interactions are shown in Supplementary Material Figs. S13 and S14.

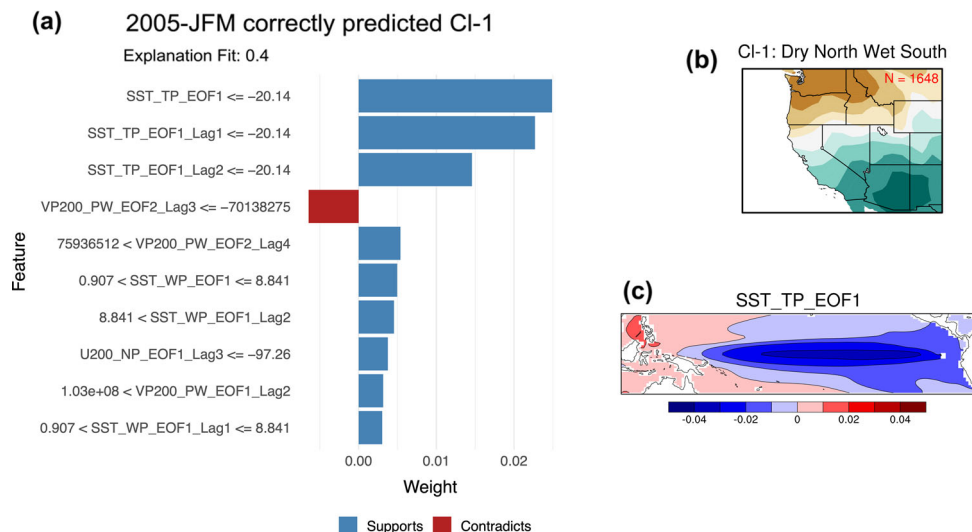
teleconnection to North America<sup>13,34</sup>. The use of velocity potential in the model allows for the representation of not only the direct influence of ENSO in the SST anomaly field but more broadly different dipole patterns of deep convection across the Pacific and Maritime continent and Indian ocean. These patterns of deep convection are known drivers of the Rossby wave response out of the tropics<sup>34</sup> such that targeting these regions through this variable appears to be an important predictor variable in the model (Fig. 5). In contrast, North Pacific SSTs and variability in the subtropical jet appear not to be as important overall predictor variables in the trained model. However, they can carry a small amount of predictive information relevant to certain clusters as described in the next section.

Moving beyond individual variable importance measures, interactions between pairs of variables are next explored in Figs. 6–7, as well as the direction of these associations. To examine which pairs of variables were among the most important in the RF model, we assessed all pairwise variable interactions across the 5000 decision trees that make up the model. These variable interactions are assessed in terms of the *mean conditional depth* of variable interactions in the decision tree, where more important variable interactions tend to occur closer to the root of the decision tree. Figure 6 shows the depth and frequency for the 25 most shallow pairwise variable interactions as determined by the mean conditional depth measure. Of these 25 variable interactions, the overall shallowest interaction was found between the first EOF of velocity potential and the first EOF of tropical Pacific SST at different lead times. The most frequent interaction was found between tropical Pacific SST in December and those in October. These specific interactions are further investigated in the following section through partial dependence plots.

Partial dependence plots quantify the direction and strength of influence of each variable on the probability of an individual cluster outcome, after accounting for the mean effects of other variables. For the overall shallowest pairwise interaction, the partial dependence plot (Fig. 7a) illustrates that La Niña-like conditions in November (positive values of SST\_TP\_EOF1\_Lag1) combined with a broad region of enhanced convection in the

Indian Ocean/Maritime continent in August (negative values of VP200\_PW\_EOF1\_Lag4) provides an overall increase in the probability of occurrence of cluster 4. For the most frequent interaction (Supplementary Material Fig. S13), the partial dependence plot shows that La Niña-like conditions in December (positive values of SST\_TP\_EOF1) and La Niña-like conditions in October (positive values of SST\_TP\_EOF1\_Lag2) combine to increase the probability of occurrence of cluster 4. The direction of this association is in close agreement with observations and previous studies, whereby La Niña conditions tend to promote a dry southwestern and wet northwestern United States. The finding that La Niña conditions preceding December further increase the strength of this association is in agreement with the physical intuition that more persistent La Niña events are more likely to produce the canonical La Niña teleconnection response. Other pairwise variable interactions can also be explored in this way to further scrutinize the model’s learned associations. For example, in Supplementary Material Fig. S14 we show the partial dependence plot for the interaction between tropical Pacific SST variability and western tropical Pacific SST variability. For this interaction, we observe that the sign of the western Pacific SST anomaly pattern can act to modulate the probability that La Niña conditions would result in widespread dry conditions over the western US. The importance of western tropical Pacific SST variability for driving ridging has been found in previous studies using both models and observations<sup>14,15</sup>.

One caveat to interpreting partial dependence plots is that they do not compute the probability of a cluster as a function of predictors while *ignoring* the effects of other predictor variables<sup>35</sup>. Instead, partial dependence plots account for the mean effect of other variables. This can make interpretation of partial dependence plots challenging in cases where multiple predictor variables are correlated with each other<sup>36</sup>. In contrast to partial dependence plots, ALE plots average and accumulate differences in the prediction across the conditional distribution to isolate the specific unbiased effects of an individual predictor of interest. Therefore, if individual variables in ALE plots show a clear relationship with the cluster outcome, one can be more confident



**Fig. 8 Explaining individual seasonal forecasts through LIME modeling.** Panel **a** shows which variables most strongly influenced the 2005-JFM correct prediction of Cluster 1 (panel **b**). Negative values of SST\_TP\_EOF and its lags (panel **c**), corresponding to a warm tropical Pacific under a weak El Niño event across October to December, most strongly favored the prediction of Cluster 1. Another LIME case study, for the incorrect 2016-JFM prediction case, is shown in Supplementary Material Fig. S17.

that certain variables do indeed provide a non-spurious unique influence on that outcome.

ALE plots for particular variables of interest are given for predicting cluster 3 (Supplementary Material Fig. S15) and cluster 4 (Fig. S16). These plots provide further evidence that, while ENSO plays a dominant role in western US precipitation predictability, other variables beyond ENSO can also provide smaller independent contributions. For example, positive values of SST\_TP\_EOF1, corresponding to La Niña conditions, most strongly increase the probability of cluster 4 occurrence, which is in agreement with the partial dependence plot from Fig. 7. The ALE plots also confirm that velocity potential at longer lead times is important (from Fig. 7), where a range of negative values for this variable slightly increases the cluster 4 occurrence. As noted earlier, a challenging component of the seasonal prediction problem is distinguishing between cluster 3 and cluster 4, related to the positioning of the anomaly dipole across the western US. Comparing the ALE plots between cluster 3 (Fig. S15) and cluster 4 (Fig. S16) elucidates what information the model has used to distinguish between these clusters when making seasonal forecasts. Since La Niña conditions favor both clusters 3 and 4, it is shown that SST variability in the western tropical Pacific and the Indian Ocean, and to a lesser extent the North Pacific are used as additional discriminants in the model. For example, while SST variability in the North Pacific was found not to be an overall important predictor earlier (Fig. 5), this can slightly increase the likelihood of cluster 3 occurrence compared to cluster 4.

**Explaining individual seasonal forecasts.** In the previous section, we provided evidence that the trained model has been capable of learning physically plausible teleconnections for western United States precipitation. In this section, we extend interpretability to go beyond that of the general model structure and towards explaining the model's decision-making process for individual seasonal forecasts. In particular, we probe what combination of factors drives the model to make an incorrect or correct seasonal forecast in a given year.

Using the LIME modeling framework, we fit simpler statistical models around the decision point of the more complex Random

Forest model. The LIME modeling framework is applied on a case-by-case basis (local interpretability), with one model fit to an individual season. We present two case studies corresponding to incorrect and correct forecasts. The 2005-JFM seasonal cluster was correctly predicted by the model, with a Dry North Wet South pattern (cluster 1). Estimates of individual predictor variables and thresholds that most strongly supported or contradicted this forecast for cluster 1 are presented in Fig. 8a. As shown, negative values of SST\_TP\_EOF1, which correspond to a warm tropical Pacific SST anomaly associated with a weak El Niño event across October to December, most strongly favored the prediction of cluster 1. Other variables, such as velocity potential at longer lead times and western tropical Pacific SST anomalies, also favored the cluster 1 forecast but these contributed less to the outcome in this case compared to ENSO.

As noted earlier, this framework can also be applied to investigate drivers of incorrect forecasts, with the forecast for 2016-JFM presented here as an example (Supplementary Material Fig. S17). In this case, the model incorrectly predicted cluster 1 to occur, whereas cluster 4 occurred. Again, weak El Niño conditions were the main driver of this incorrect prediction. However, patterns of SST variability in the western tropical Pacific and the Indian Ocean presented a tug-of-war to contradict the prediction but with overall smaller weights. This incorrect prediction was not unique to the Random Forest model, with all other machine learning models and NMME models analyzed making similar incorrect forecasts. Various other studies have concluded that this particular event was largely dominated by less predictable atmospheric variability, making it difficult to predict on seasonal timescales<sup>4,11,12</sup>.

It is important to note that the LIME framework is only an approximate estimate of the machine learning model's more complex decision-making process at that locality. In the two cases presented, the simple model was found to provide a reasonably good approximation ( $R^2 = 0.4-0.5$ ) of the model complexity, but less explainable cases can also be found. While acknowledging this caveat, we suggest that the ability to explain individual forecasts in this way could be very useful. In dynamical models used for seasonal forecasts, it is often difficult to formally quantify



what boundary conditions have most heavily influenced an individual forecast without running further resource-expensive diagnostic experiments. Subsequently, in practice, it often takes considerable time after the event for diagnostic studies to reevaluate and quantify the physical drivers of a forecast, and interpretation from this can be inconclusive. In contrast, here we have illustrated how local interpretable machine learning can provide plausible explanations for what variables contributed most strongly to a particular forecast outcome at a negligible computational cost. In practice, the main advantage of this is that these local interpretability plots (Fig. 8) can rapidly be produced and presented alongside the seasonal forecast in real-time.

**Implications and future directions.** The machine learning approach to seasonal forecasting tested here shows promise both in terms of competitive accuracy and in terms of ability to learn physically plausible teleconnections. The proposed interpretable machine learning approach could also be applied more broadly in future work to better understand and compare teleconnections between different climate models, as well as assessing the possibility of non-stationarity in certain teleconnections due to climate change.

A number of different pathways may exist for further improvements in seasonal forecast skills. One major advantage is that training machine learning models on climate model simulations can leverage substantial existing investments in large climate model ensembles. Indeed, the number of modeling groups performing large initial condition model experiments has increased considerably in recent years<sup>37</sup>, providing further opportunities to train on different climate models and better understand different structural uncertainties that contribute to seasonal forecast uncertainty. This large and growing set of model simulations available for training is in contrast to the traditional approach of training on observational data, where only a single additional training sample becomes available each year. In future work, we plan to assess potential skill improvement from training on certain climate model simulations that have a reasonably well resolved quasi-biennial oscillation (QBO), which contains a large amount of memory and has been highlighted recently as an important source of predictability for North American precipitation<sup>18</sup>.

It is notable that the Random Forest, one of the computationally simpler machine learning models tested, ranked as one of the top-performing models. This carries a practical advantage since the Random Forest is typically more readily interpretable (as explored in Figs. 5–7) and has fewer tunable parameters with relatively little sensitivity to these parameter choices. For the LSTM model, a simple implementation (single layer LSTM followed by 20 neurons in a single layer) was found to achieve competitive results compared to more complex architectures. However, future work may explore different LSTM implementations, including testing deeper LSTM architectures and bidirectional LSTM<sup>38</sup>. Transfer learning may be another approach to further improve skill<sup>25</sup>. Transfer learning in this context involves generating the main associations and weights from training on large climate model simulations then updating these pretrained weights on a separate set of observations.

## Summary

This study has tested a novel approach for seasonal forecasting western US precipitation. In particular, a range of machine learning approaches have been trained on large climate model simulations, and their predictions combined in an ensemble to predict large-scale patterns of precipitation anomalies. The main findings from this study are:

Classification accuracy is generally higher in JFM compared to NDJ seasons. In both seasons, the machine learning models display skillful predictions relative to baselines, and can

compete with or out-compete dynamical forecast models from NMME.

The widespread wet pattern of precipitation (cluster 2) was consistently the most difficult pattern to forecast in all machine learning models. Post-processing, where different clusters are combined, can increase the accuracy further (accuracy: 70–80%) but comes at the cost of providing a less precise forecast (i.e., larger spatial smoothing).

Focusing on the Random Forest model, we have investigated both global and local interpretability. In terms of global interpretability, as expected, ENSO is the dominant source of seasonal predictability. Other variables, namely velocity potential anomalies across the Indian Ocean and Maritime Continent, and SST anomalies across the western tropical Pacific can modulate the probability that ENSO will result in a certain precipitation cluster. The interpretability results provide confidence that the model is capable of learning physically plausible teleconnections from a large pool of candidate predictor variables.

Local interpretability provides estimates of what variables have influenced an individual seasonal forecast. Examples of local interpretability are presented showing how, for specific seasons, conditions beyond ENSO have influenced a specific forecast. Presenting local interpretability plots alongside the seasonal forecast may help build trust in the predictions from machine learning.

We suggest that this approach to seasonal forecasting offers a promising path forward. Compared to the traditional approach of training statistical models on observational data, the large sample size enabled by training on large climate model simulations helps overcome sampling issues and allows for nonlinear interactions to be represented. Further skill improvements may come from training machine learning models on multiple climate models through the same framework.

## Methods

### Overview of the framework for machine learning with large ensemble climate simulations

This section provides an overview of the framework implemented here for machine learning-based predictions from large ensemble climate simulations. Figure 1 outlines this methodology, showing oceanic (Fig. 1a) and atmospheric (Fig. 1b) predictor variables and regions from the CESM-LENS model, described in Section “Climate model training data” below. The predictor variables are based on applying EOF analysis to different regions and variables (described in Section “Machine learning predictor variables”). Using these EOF-derived predictor variables in model training, the four machine learning models tested in this study are shown in Fig. 1c. The predictand variable (Fig. 1d) targets widespread spatial patterns of precipitation derived from applying K-means clustering to CESM-LENS precipitation data (Section “Machine learning predictand variable”). Predicting the occurrence of these larger-scale precipitation features (Fig. 1d) has the advantage that these spatial scales are well aligned with the typical area of ridges and troughs along Rossby wave trains that form an important source of seasonal predictability in this region<sup>14,15,39,40</sup>. After training and calibrating each of the four machine learning models on CESM-LENS data through this framework, the same models are then forced by observational and reanalysis data (described in Section “Observational and reanalysis data”) and used to make out-of-sample seasonal forecasts for NDJ and JFM seasons across the observed record (1980–2020).

**Climate model training data.** The limited record length of observational data at the seasonal time resolution leads us to explore the use of climate model data when training various machine learning models. For this purpose, we use simulations from the CESM-LENS single-model large ensemble<sup>41,37</sup>, comprising 40 ensemble members spanning years 1920–2005 with historical forcing from the fifth phase of the Coupled Model Intercomparison Project (CMIP5) design protocol. The CESM-LENS uses the Community Earth System Model v1 (CESM1), with the Community Atmosphere Model (CAM) v5, run at approximately 1° resolution with fully coupled atmosphere, ocean, land, and sea-ice components. Each of the CESM-

LENS ensemble members represents a physically plausible and unique trajectory of the climate system (e.g., different phases of low-frequency variability will occur at different times across the historical record) solely due to internally generated climate variability. Data from CESM-LENS used for training machine learning models were as follows: SST, zonal and meridional wind at 200 hPa (U200, V200), velocity potential at 200 hPa (VP200), geopotential height at 500 hPa (Z500), and total precipitation.

A number of studies have evaluated the performance of CESM1 in terms of simulating low-frequency variability in the tropics (i.e., including ENSO) and related teleconnections into the North Pacific<sup>42–45</sup>. These considerations are relevant to training machine learning on CESM1, as systematic biases in teleconnections also have the potential to be learned during training. A primary focus of CESM1 model development was to improve ENSO variability and teleconnections as these were known to have large deficiencies across previous model versions (e.g., CCSM3)<sup>45</sup>. The much-improved fidelity of ENSO variability and teleconnections across successive versions of the model was largely the result of targeted changes to the atmospheric deep convection parameterization<sup>46,47</sup>. CESM1 has shown to generally perform well in terms of temporal characteristics including the asymmetry of El Niño and La Niña duration<sup>42</sup>. However, the amplitude and variability of ENSO events are both known to be larger than that observed across the 20th century<sup>42</sup>.

**Machine learning predictor variables.** All predictor variables first underwent dimension reduction through EOF analysis. The purpose of this was to: (1) isolate dominant spatial modes of variability and (2) reduce the amount of potentially redundant data by transferring from gridded data to a smaller number of principal components more manageable in model training. A similar approach is typically implemented as a first step in more traditional CCA for seasonal forecasting. All predictor variables were based on monthly mean values. For each predictor variable, the first four EOFs were retained which collectively explain at least 50% of its variance. The spatial patterns of EOFs and the percent variance explained are shown in Supplementary Material Figs. S2–S8.

As detailed in Fig. 2a, EOF-derived predictor variables for SST were chosen to target the following regions: tropical Pacific (TP), western tropical Pacific (WP), Indian Ocean (IO), and North Pacific (NP). These regions were targeted based on previous studies (see Section “Sources of seasonal predictability”) indicating plausible physical teleconnections to western US precipitation. EOF-derived predictor variables for atmospheric circulation (Fig. 1b) target velocity potential anomalies at 200 hPa across the wider tropical Pacific (PW) and the Indian Ocean, zonal wind anomalies at 200 hPa across the North Pacific (NP), and geopotential height anomalies at 500 hPa across the Eastern North Pacific (ENP). The velocity potential field targets broad spatial patterns of anomalous deep convection that drive a Rossby wave response in the extratropics. North Pacific subtropical jet variability was also included because of the waveguiding influence on tropical–extratropical teleconnections<sup>48–50</sup> as well as the relevance to western US precipitation through more localized jet regimes over the northeast Pacific<sup>11,51</sup>. Other variables considered included tropical and high-latitude stratospheric variability, including the QBO and sudden stratospheric warming events. However, stratospheric variability is not well resolved in this low-top model version of CESM1<sup>52</sup>, and as expected, sensitivity testing found that these variables did not add additional skill and were subsequently removed.

**Machine learning predictand variable.** The predictand variable is derived from K-means clustering of standardized seasonal (3-monthly) precipitation anomalies over the western US for two separate seasons: NDJ and JFM. Cluster analysis was used to isolate recurrent large-scale features of precipitation variability in this region (Fig. 1d). K-means clustering requires the user to select a specific number of clusters, which we set as four clusters for the following reasons. Previous research has highlighted that the first two modes of seasonal precipitation anomalies in this region explain approximately 60% of the variance<sup>39</sup>. The first mode is associated with widespread wet/dry conditions across the entire region, while the second mode is associated with a north–south dipole of precipitation anomalies<sup>53</sup>. The first four clusters from K-means (Fig. 1d) are a very close match to these dominant modes of seasonal precipitation. Furthermore, by choosing four clusters, the clusters trained from CESM-LENS (shown in Fig. 1d) are a close match to those trained from observational data (Supplementary Material Fig. S1). While including more than four clusters can provide more regional detail in precipitation, the prediction accuracy from the machine learning methods tested was found to decrease when additional clusters were added in sensitivity testing. This suggests that the four clusters extracted from K-means broadly represent the main predictable components of precipitation in CESM-LENS on seasonal timescales.

**Observational and reanalysis data.** After training on CESM-LENS, the trained machine learning models were taken offline and tested for making out-of-sample predictions on observations (years 1980–2020). For this purpose, the same set of observed predictor variables are needed as those used in model training. Specifically, SST data were obtained from ERSSTv5<sup>54</sup>, all atmospheric circulation fields were from ERA5<sup>55</sup>, and the total precipitation was from CPC-Unifed at 0.25°

resolution<sup>56</sup>. The dimension reduction of these predictor and predictand variables was applied to both CESM-LENS and observations.

**Random Forests.** Random Forests (RF)<sup>57</sup> is a supervised machine learning algorithm consisting of an ensemble of decision trees. Different decision trees are developed by taking random subsets of predictor variables and data cases, which reduces the correlation between individual trees. The purpose of using multiple decision trees is that the variance in the prediction is reduced compared to predictions from individual trees that are often prone to overfitting on the training data. Each tree is built using a bootstrapped sample, and records that are not used in building the decision tree are referred to as the out-of-bag sample. In a number of settings on tabular data, RFs are often shown to be capable of producing similar classification accuracy compared to more complex machine learning methods, while retaining a somewhat higher degree of interpretability. Another advantage of RFs is the relatively small number of parameters required in model tuning and the relative insensitivity to these choices<sup>57</sup>.

Two parameters were tuned in the RF model: the number of trees set to 5000, and the number of variables randomly sampled at each split set to 10. These parameter choices were based on tuning across the CESM-LENS training dataset, though sensitivity testing revealed stable results across a number of parameter choices provided that the number of trees was sufficiently large. For the RF training, predictor variables were lagged based on the memory of each predictor variable (Supplementary Material Figs. S9 and 10) and through sensitivity testing to the of out-of-bag accuracy across the CESM-LENS training to how each variable was lagged. The first EOF of each SST variable/region was lagged at 1-month intervals up to 12 months, and the second EOFs were lagged up to 6 months. The exception was for SST in the North Pacific, which was not lagged, as we found very little sensitivity to adding additional lags. The first EOF of U200 was also lagged up to 6-months, given the memory of this variable. All other predictor variables were not lagged, such that only October values were used to make the NDJ predictions and only December values were used to make the JFM predictions.

**XGBoost.** Extreme gradient boosting (XGBoost) is a recent implementation of gradient boosted decision trees for supervised machine learning<sup>58</sup>. XGBoost relies on the concept of boosting—that multiple “weak” learners (i.e., underfit to the data) can be more effectively combined to produce a single “strong” learner. The training proceeds by iteratively growing individual decision trees that target misclassifications from the previous weak learners, giving them additional weight across subsequent training iterations. Recently, XGBoost has been a consistent top performer in terms of classification accuracy for tabular datasets across an extensive range of applied machine learning problems<sup>58</sup>.

Compared to RF, XGBoost requires a larger number of parameters to be tuned in the model training. Among the most important parameters are the number of rounds for boosting (*nrounds*), how deep the trees can grow (*max\_depth*), the learning rate to control how conservative the boosting is performed (*eta*), and the minimum loss reduction required to make a further tree partition (*gamma*). These parameter values (Supplementary Material Table S1) were tuned based on multiclass classification accuracy in the CESM-LENS validation dataset from a random search across a range of values, performed separately for models predicting JFM and NDJ seasons. For tuning purposes, we split the CESM-LENS dataset as 80%/20% for training and validation, respectively with the same variables and lags were used as in the RF model, described in Section “Random Forests”.

**Neural networks.** Neural networks (NN) approximate nonlinear functions and processes<sup>59</sup> through a series of feed-forward matrix operations. NNs pass predictor input variables through a series of hidden layers, to a specified output layer. Each layer is described by the number of nodal points in that layer with the initial layer being the number of input variables. Nodes from adjacent model layers are connected via model weights. The hidden nodal point values are determined by the sum of the product of associated model weights and the input values from the previous layer. Each nodal point is then “activated” by a nonlinear function before passing the variables to the following layer. The task of training a NN is to learn the optimal nodal weights, computed iteratively through backward optimization and gradient descent. In particular, each iteration seeks to minimize the cost of a specified loss function, by determining the gradient field of the weights and taking a small step in the direction opposite this gradient. The series of multiple hidden layers, and the optimization process, gives rise to the term Deep Learning.

A Deep Feed Forward NN was implemented in which all nodes are fully connected, without enforcing sparsity. The final architecture and parameters were selected through a hyperparameter search, with the minimum error on the validation data set (20% of the CESM-LENS dataset) used to determine the final network parameters and architecture (Supplementary Material Table S2). The NN utilizes an Adam optimizer<sup>60</sup>, the Rectified Linear Unit (ReLU) activation function, a 0.001 learning rate, a batch size of 100, dropout regularization, and a categorical cross-entropy loss. The network was trained with 50 epochs and saved whenever validation accuracy improves. All predictor variables (Section 2.3) were lagged by 12 months and used in the final model. Class imbalances were accounted for by applying a scalar value that weights the cross-entropy loss function during training proportionally to categorical representation in the training dataset.

**LSTM networks.** Long short-term memory (LSTM) networks are a type of recurrent NN that can learn dependence in sequential data<sup>61,62</sup>. LSTM networks have an internal state that aims to model information about past observations (inputs) for a variable number of steps in the sequence. The advantage of this approach, compared to traditional feed-forward NNs, is that persistence and memory of past information are explicitly modeled in the network and used when making predictions. In the context of seasonal forecasting, this feature of LSTMs is valuable since a sequence of past events (e.g., the gradual development and persistence of ENSO over several months), as opposed to an individual data point/month, is likely provides additional predictive information.

All predictor variables were used in the final model, but here the sequence length history (i.e., how far the model looks back at past events) was treated as a hyperparameter. Other hyperparameters were the number of LSTM neurons, training epochs, and batch size, which were tuned across the validation data set to determine the final network values (Supplementary Material Table S3). The network architecture used for both JFM and NDJ models was fixed with the LSTM layer followed by 20 neurons in a single layer. The LSTM network used an Adam optimizer<sup>60</sup> with categorical cross-entropy loss. Dropout regularization was implemented to reduce overfitting, where the network nodes are probabilistically dropped out of weight updates during model training. Further details on the training data and software implementation for each machine learning model are presented in Supplementary Material Table S4 and Table S5.

**Interpretable machine learning.** The goal of interpretable machine learning is to quantify which predictor variables are overall most influential in the model (global interpretation), as well as estimating how an individual classification/forecast was made (local interpretation). In recent years, applied machine learning research has focused heavily on developing these techniques, which are directly relevant to applications in atmospheric science<sup>63–65</sup>. Here, we describe the implementation of three separate approaches to target global and local interpretability, specifically from the RF model.

First, to target global interpretation, we explore the most important individual predictor variables used for making seasonal predictions. Three metrics were considered as follows: relative mean decrease accuracy, mean minimum tree depth, and root. Relative mean decrease accuracy quantifies the decrease in classification accuracy from shuffling individual predictor variables. If shuffling results in a relatively large decrease in the test set accuracy (i.e., larger errors across the out-of-bag samples) then the variable is seen to be important since shuffling has now broken a previously important relationship. Mean minimum tree depth quantifies the average depth of a particular predictor variable across multiple decision trees. Since more consequential variables are positioned closer to the root of the decision tree, a lower minimum tree depth value signifies larger variable importance. Similarly, the root metric measures variable importance by counting the number of times a particular predictor variable is positioned at the root of a decision tree.

Second, also targeting global interpretation, we explore the most important predictor variable interactions between all possible pairs of predictor variables. To do so, across all decision trees, we compute the mean conditional depth of all pairwise interactions in terms of their frequency of interaction occurrence and their depth of interaction occurrence<sup>66,67</sup>. This enables analysis of variable interaction importance since variables that interact closer to the root of the tree and with a higher frequency will be more consequential overall in determining the prediction. Partial dependence plots compute the probability of an outcome as a function of two predictor variables after accounting for the average effects of all other variables<sup>63,68</sup>. This analysis allows us to examine the average direction and strength of influence for key predictor variables, as well as examining the potential influence of nonlinear interactions on the probability of a particular outcome. Accumulated local effects plots (ALE plots)<sup>36</sup> are also computed, which can be considered an extension of partial dependence plots. ALE plots are designed to limit potential issues associated with the multicollinearity of predictor variables that can make partial dependence plots challenging to interpret.

Third, we implement the local interpretable model-agnostic explanations modeling framework (LIME)<sup>69</sup> to explain why a particular cluster was forecasted on a particular target date (i.e., local interpretation). In the context of seasonal forecasting, LIME is used to explain and rank which predictor variables were used by the model to make a particular forecast. The LIME modeling framework aims to explain the importance of predictor variables by finding a simple linear solution that approximates the original model's decision function near to a particular case. Through perturbing input variables across the model's nonlinear decision function, this much simpler linear model is fit to explain how the more complex RF model behaves locally.

**Comparisons to dynamical forecast models.** The output of a number of dynamical models from the North American multi-model ensemble (NMME) phase 2 models<sup>19,20</sup> were analyzed to compare skill with the machine learning-based models (Supplementary Material Table S6). In particular, the 3-month forecasted standardized seasonal precipitation anomaly from each model's ensemble mean was projected onto the K-means clusters. This projection into cluster-space allowed direct comparisons between the machine learning-based models and the dynamical models. Forecasts for November through January (NDJ) were initialized in October, and forecasts for January through March (JFM) were initialized in December, with all available hindcast years used in each model.

## Data availability

All data used in this study are publicly available. CESM-LENS 40-member ensemble data were provided by the Climate Data Gateway at NCAR: <https://www.cesm.ucar.edu/projects/community-projects/LENS/data-sets.html> NMME data were provided by the IRI Data Library: <http://iridl.ldeo.columbia.edu/SOURCES/Models/NMME/ERA5> data were provided by the Copernicus Climate Data Store: <https://cds.climate.copernicus.eu/#/search?text=ERA5&type=dataset> ERSSTv5 data were provided by NOAA/OAR/ESRL PSL: <https://psl.noaa.gov/data/gridded/data.noaa.ersst.v5.html> CPC US Unified Precipitation data were provided by NOAA/OAR/ESRL PS: <https://psl.noaa.gov/data/gridded/data.unified.daily.conus.html>

## Code availability

All code used in this study is available from the corresponding author upon request.

Received: 8 March 2021; Accepted: 23 June 2021;

Published online: 10 August 2021

## References

1. Dettinger, M. D., Ralph, F. M., Das, T., Neiman, P. J. & Cayan, D. R. Atmospheric rivers, floods and the water resources of California. *Water* **3**, 445–478 (2011).
2. DeFlorio, M. J., Waliser, D. E., Guan, B., Ralph, F. M. & Vitart, F. Global evaluation of atmospheric river subseasonal prediction skill. *Clim. Dyn.* **52**, 3039–3060 (2019).
3. DeFlorio, M. J. et al. Experimental subseasonal-to-seasonal (S2S) forecasting of atmospheric rivers over the Western United States. *J. Geophys. Res.* **124**, 11242–11265 (2019).
4. Kumar, A. & Chen, M. What is the variability in US west coast winter precipitation during strong El Niño events? *Clim. Dyn.* **49**, 2789–2802 (2017).
5. Howitt, R., Medellín-Azuara, J., MacEwan, D., Lund, J. R., & Sumner, D. *Economic Analysis of the 2014 Drought for California Agriculture*. (Center for Watershed Sciences University of California, Davis, CA., 2014).
6. Lund, J., Medellín-Azuara, J., Durand, J. & Stone, K. Lessons from California's 2012–2016 drought. *J. Water Resour. Plan. Manag.* **144**, 04018067 (2018).
7. Barnston, A. G. & Smith, T. M. Specification and prediction of global surface temperature and precipitation from global SST using CCA. *J. Clim.* **9**, 2660–2697 (1996).
8. Trenberth, K. E. et al. Progress during TOGA in understanding and modeling global teleconnections associated with tropical sea surface temperatures. *J. Geophys. Res.* **103**, 14291–14324 (1998).
9. Gibson, P. B., Waliser, D. E. & DeFlorio, M. J. A critical examination of a newly proposed interhemispheric teleconnection to Southwestern US winter precipitation. *Nat. Commun.* **10**, 2687 (2019).
10. Chapman, W. E. et al. Monthly modulations of ENSO teleconnections: implications for potential predictability in North America. *J. Clim.* <https://doi.org/10.1175/jcli-d-20-0391.1> (2021).
11. Wang, S., Anichowski, A., Tippett, M. K. & Sobel, A. H. Seasonal noise versus subseasonal signal: forecasts of California precipitation during the unusual winters of 2015–2016 and 2016–2017. *Geophys. Res. Lett.* **44**, 9513–9520 (2017).
12. Cash, B. A. & Burls, N. J. Predictable and unpredictable aspects of U.S. West Coast Rainfall and El Niño: understanding the 2015/16 event. *J. Clim.* **32**, 2843–2868 (2019).
13. Patricola, C. M. et al. Maximizing ENSO as a source of western US hydroclimate predictability. *Clim. Dyn.* **54**, 351–372 (2020).
14. Teng, H. & Branstator, G. Causes of extreme ridges that induce California droughts. *J. Clim.* **30**, 1477–1492 (2017).
15. Gibson, P. B. et al. Ridging associated with drought across the Western and Southwestern United States: characteristics, trends, and predictability sources. *J. Clim.* **33**, 2485–2508 (2020).
16. Siler, N., Kosaka, Y., Xie, S.-P. & Li, X. Tropical ocean contributions to California's surprisingly Dry El Niño of 2015/16. *J. Clim.* **30**, 10067–10079 (2017).
17. Kidston, J. et al. Stratospheric influence on tropospheric jet streams, storm tracks and surface weather. *Nat. Geosci.* **8**, 433–440 (2015).
18. Mariotti, A. et al. Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond. *Bull. Am. Meteorol. Soc.* **101**, E608–E625 (2020).
19. Kirtman, B. P. et al. The North American Multimodel Ensemble: phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. *Bull. Am. Meteorol. Soc.* **95**, 585–601 (2014).
20. Becker, E., Kirtman, B. P. & Pegion, K. Evolution of the North American multi-model ensemble. *Geophys. Res. Lett.* **47**, e2020GL087408 (2020).
21. Slater, L. J., Villarini, G. & Bradley, A. A. Evaluation of the skill of North-American Multi-Model Ensemble (NMME) Global Climate Models in predicting average and extreme precipitation and temperature over the continental USA. *Clim. Dyn.* **53**, 7381–7396 (2019).

22. Barnston, A. G. et al. Long-lead seasonal forecasts—where do we stand? *Bull. Am. Meteorol. Soc.* **75**, 2097–2114 (1994).
23. Witten, D. M. & Tibshirani, R. J. Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat. Appl. Genet. Mol. Biol.* <https://doi.org/10.2202/1544-6115.1470> (2009).
24. Cohen, J. et al. S2S reboot: an argument for greater inclusion of machine learning in subseasonal to seasonal forecasts. *WIREs Clim. Change* **10**, e00567 (2019).
25. Ham, Y.-G., Kim, J.-H. & Luo, J.-J. Deep learning for multi-year ENSO forecasts. *Nature* **573**, 568–572 (2019).
26. van der Ploeg, T., Austin, P. C. & Steyerberg, E. W. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med. Res. Methodol.* **14**, 137 (2014).
27. DelSole, T. & Banerjee, A. Statistical seasonal prediction based on regularized regression. *J. Clim.* **30**, 1345–1361 (2017).
28. Stevens, A. et al. Graph-guided regularized regression of Pacific Ocean climate variables to increase predictive skill of southwestern U.S. winter precipitation. *J. Clim.* **34**, 737–754 (2021).
29. Chen, L.-C., van den Dool, H., Becker, E. & Zhang, Q. ENSO precipitation and temperature forecasts in the North American multimodel ensemble: composite analysis and validation. *J. Clim.* **30**, 1103–1125 (2017).
30. Gong, X., Barnston, A. G. & Ward, M. N. The effect of spatial aggregation on the skill of seasonal precipitation forecasts. *J. Clim.* **16**, 3059–3071 (2003).
31. Huang, B., Shin, C.-S. & Kumar, A. Predictive skill and predictable patterns of the U.S. seasonal precipitation in CFSv2 reforecasts of 60 years (1958–2017). *J. Clim.* **32**, 8603–8637 (2019).
32. van Straaten, C., Whan, K., Coumou, D., van den Hurk, B. & Schmeits, M. The influence of aggregation and statistical post-processing on the subseasonal predictability of European temperatures. *Q. J. R. Meteorol. Soc.* **146**, 2654–2670 (2020).
33. Lapuschkin, S. et al. Unmasking Clever Hans predictors and assessing what machines really learn. *Nat. Commun.* **10**, 1096 (2019).
34. Molteni, F., Stockdale, T. N. & Vitart, F. Understanding and modelling extra-tropical teleconnections with the Indo-Pacific region during the northern winter. *Clim. Dyn.* **45**, 3119–3140 (2015).
35. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. (Springer Science & Business Media, 2009).
36. Apley, D. W. & Zhu, J. Visualizing the effects of predictor variables in black box supervised learning models. *J. R. Stat. Soc.* **82**, 1059–1086 (2020).
37. Deser, C. et al. Insights from Earth system model initial-condition large ensembles and future prospects. *Nat. Clim. Change* **10**, 277–286 (2020).
38. Schuster, M. & Paliwal, K. K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **45**, 2673–2681 (1997).
39. Cook, B. I. et al. Revisiting the leading drivers of Pacific Coastal drought variability in the contiguous United States. *J. Clim.* **31**, 25–43 (2018).
40. Gibson, P. B. et al. Subseasonal-to-seasonal hindcast skill assessment of ridging events related to drought over the Western United States. *J. Geophys. Res.* **125**, e2020JD033655 (2020).
41. Kay, J. E. 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. *Bull. Am. Meteorol. Soc.* **96**, 1333–1349 (2015).
42. DiNezio, P. N., Deser, C., Okumura, Y. & Karspeck, A. Predictability of 2-year La Niña events in a coupled general circulation model. *Clim. Dyn.* **49**, 4237–4261 (2017).
43. Deser, C., Simpson, I. R., McKinnon, K. A. & Phillips, A. S. The Northern Hemisphere Extratropical Atmospheric Circulation Response to ENSO: how well do we know it and how do we evaluate models accordingly? *J. Clim.* **30**, 5059–5082 (2017).
44. Swain, D. L., Langenbrunner, B., Neelin, J. D. & Hall, A. Increasing precipitation volatility in twenty-first-century California. *Nat. Clim. Change* **8**, 427–433 (2018).
45. Danabasoglu, G. et al. The Community Earth System Model Version 2 (CESM2). *J. Adv. Model. Earth Syst.* **12**, e2019MS001916 (2020).
46. Neale, R. B., Richter, J. H. & Jochum, M. The impact of convection on ENSO: from a delayed oscillator to a series of events. *J. Clim.* **21**, 5904–5924 (2008).
47. Deser, C. et al. ENSO and Pacific decadal variability in the community climate system model version 4. *J. Clim.* **25**, 2622–2651 (2012).
48. Branstator, G. Circumglobal teleconnections, the jet stream waveguide, and the North Atlantic Oscillation. *J. Clim.* **15**, 1893–1910 (2002).
49. Seo, K.-H. & Lee, H.-J. Mechanisms for a PNA-Like teleconnection pattern in response to the MJO. *J. Atmos. Sci.* **74**, 1767–1781 (2017).
50. Wang, J. et al. MJO teleconnections over the PNA region in climate models. Part II: impacts of the MJO and basic state. *J. Clim.* **33**, 5081–5101 (2020).
51. Neelin, J. D., Langenbrunner, B., Meyerson, J. E., Hall, A. & Berg, N. California winter precipitation change under Global Warming in the Coupled Model Intercomparison Project Phase 5 Ensemble. *J. Clim.* **26**, 6238–6256 (2013).
52. Richter, J. H. et al. Progress in simulating the quasi-biennial oscillation in CMIP models. *J. Geophys. Res.* **125**, e2019JD032362 (2020).
53. DeFlorio, M. J., Pierce, D. W., Cayan, D. R. & Miller, A. J. Western U.S. extreme precipitation events and their relation to ENSO and PDO in CCSM4. *J. Clim.* **26**, 4231–4243 (2013).
54. Huang, B. et al. Extended reconstructed sea surface temperature, Version 5 (ERSSTv5): upgrades, validations, and intercomparisons. *J. Clim.* **30**, 8179–8205 (2017).
55. Hersbach, H. et al. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **146**, 1999–2049 (2020).
56. Higgins, R. W., Shi, W., Yarosh, E., & Joyce, R. Improved United States precipitation quality control system and analysis. NCEP/Climate Prediction Center Atlas 7 (2000).
57. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
58. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, 2016).
59. Nielsen, M. A. *Neural Networks and Deep Learning*. (Determination Press, San Francisco, CA, 2015).
60. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. Preprint at [arXiv https://arxiv.org/abs/1412.6980](https://arxiv.org/abs/1412.6980) (2014).
61. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
62. Gers, F. A. & Schmidhuber, J. in *Proc. IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*. 189–194; vol.183.
63. McGovern, A. et al. Making the black box more transparent: understanding the physical implications of machine learning. *Bull. Am. Meteorol. Soc.* **100**, 2175–2199 (2019).
64. Barnes, E. A. et al. Indicator patterns of forced change learned by an artificial neural network. *J. Adv. Model. Earth Syst.* <https://doi.org/10.1029/2020MS002195> (2020).
65. Toms, B. A., Barnes, E. A. & Ebert-Uphoff, I. Physically interpretable neural networks for the geosciences: applications to earth system variability. *J. Adv. Model. Earth Syst.* **12**, e2019MS002002 (2020).
66. Ishwaran, H., Kogalur, U. B., Gorodeski, E. Z., Minn, A. J. & Lauer, M. S. High-dimensional variable selection for survival data. *J. Am. Stat. Assoc.* **105**, 205–217 (2010).
67. Paluszynska, A. *Structure Mining and Knowledge Extraction from Random Forest with Applications to the Cancer Genome Atlas Project*, Master's thesis. (University of Warsaw, Warsaw, 2017).
68. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
69. Ribeiro, M. T., Singh, S. & Guestrin, C. in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1135–1144 (Association for Computing Machinery, San Francisco, California, USA, 2016).

## Acknowledgements

Authors P.B.G., W.E.C., L.D.M., and M.J.D. acknowledge funding and support from the California Department of Water Resources Atmospheric River Program (grant No. 4600010378 TO#15 Am 22). Authors A.A. and D.E.W. acknowledge funding and support from the California Department of Water Resources (JPL/Caltech Task #82-19834). A.A. and D.E.W.'s contribution to this research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with NASA.

## Author contributions

Authors P.B.G., W.E.C., and A.A. carried out the tuning, training, and validation of the models, and author P.B.G. prepared the figures. Authors P.B.G., W.E.C., A.A., L.D.M., M.J.D., and D.E.W. were involved substantially in the conception and design of the study and writing of the paper.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s43247-021-00225-4>.

**Correspondence** and requests for materials should be addressed to P.B.G.

**Peer review information** *Communications Earth and Environment* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: Heike Langenberg.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021