SCIENTIFIC REPORTS
nature research

OPEN

Long-lead Prediction of ENSO Modoki Index using Machine Learning algorithms

Manali Pal[1], Rajib Maity[1*], J. V. Ratnam[2], Masami Nonaka[2] & Swadhin K. Behera[2]

The focus of this study is to evaluate the efficacy of Machine Learning (ML) algorithms in the long-lead prediction of El Niño (La Niña) Modoki (ENSO Modoki) index (EMI). We evaluated two widely used non-linear ML algorithms namely Support Vector Regression (SVR) and Random Forest (RF) to forecast the EMI at various lead times, viz. 6, 12, 18 and 24 months. The predictors for the EMI are identified using Kendall's tau correlation coefficient between the monthly EMI index and the monthly anomalies of the slowly varying climate variables such as sea surface temperature (SST), sea surface height (SSH) and soil moisture content (SMC). The importance of each of the predictors is evaluated using the Supervised Principal Component Analysis (SPCA). The results indicate both SVR and RF to be capable of forecasting the phase of the EMI realistically at both 6-months and 12-months lead times though the amplitude of the EMI is underestimated for the strong events. The analysis also indicates the SVR to perform better than the RF method in forecasting the EMI.

The El Niño (La Niña) Modoki (ENSO Modoki, hereafter EM)[1] is a newly acknowledged phenomenon characterized by warm (cool) central Pacific sea surface temperature (SST) flanked by cool (warm) eastern and western Pacific SSTs. The EM events affect the global climate at various time scales. The EM affects the equatorial or near equatorial countries by the modified Walker circulation with rising (sinking) motion in the central equatorial Pacific and sinking (rising) motion over the west and east Pacific during EM warm (cold) events and other parts of the globe are affected by the atmospheric teleconnections due to the distribution of heating associated with the equatorial SST anomalies during the EM events[2–8].

Although impacts of EM events have been well established, the EM is apparently not so well predicted at long lead times by current operational climate forecast models[2,9–14]. The Bureau of Meteorology Predictive Ocean Atmosphere Model for Australia (POAMA) coupled seasonal forecast model showed a partial success in predicting differences between Modoki and canonical El Niños one season ahead with correlation coefficient more than 0.6[15,16]. APEC Climate Center (APCC) Multi-Model Ensemble (MME) seasonal forecast system shows the ability to predict the patterns of tropical Pacific SST anomaly (SSTA) of the Modoki events four months ahead with a high correlation coefficient i.e. 0.8[17]. However, the predictability of anomalous SST patterns in the APCC MME is seasonally dependent. The IAP-DecPreS near-term climate prediction system, though could predict the EMI with a good skill (correlations coefficient of 0.62 and 0.53) at 4 and 7 months lead, has limited skill (correlation coefficient 0.43) at a lead time of 10 months and beyond[18]. All the above models have difficulties in forecasting the amplitude of the EMI events though they forecast the phase of the EMI realistically. The limited skill in predicting the ENSO Modoki index (EMI) in terms of long lead times by the current seasonal forecasting systems, on the face of huge benefit in predicting it, motivated us to look for alternative methods to forecast the EMI.

Statistical EM prediction based on the non-linear Machine Learning (ML) algorithms could be a potential alternative to the dynamical model based prediction. The ML based prediction has generally shown good skill in forecasting events, though the method has limited capability to understand the underlying processes[19]. The skill of the ML algorithms stems from the use of observed data for the training. The ML algorithms, unlike climate models, are less computationally intensive. Two widely used ML algorithms namely Support Vector Regression (SVR), and Random Forest (RF) are used here.

The RF technique proposed by Breiman (2001)[20] is popular for classification, prediction, studying variable importance, variable selection, and outlier detection. It consists of an ensemble of simple tree predictors where

[1]Department of Civil Engineering, Indian Institute of Technology Kharagpur, Kharagpur, 721302, West Bengal, India. [2]Application Laboratory, Japan Agency for Marine-Earth Science and Technology, Yokohama, Japan. *email: rajib@civil.iitkgp.ac.in

each tree yields a response presented with a set of predictor values. In regression problems, the responses are averaged to estimate the dependent variable. The SVR[21] minimizes the expected error of a learning machine thus reducing the problem of overfitting. This is a robust and proficient technique for both classification and regression. A lot of studies on the applications of RF and SVR can be found where these techniques outperformed or performed with same skill as other established techniques. The studies showing the skills of these two algorithms include a study on drought forecast to predict the time series of monthly standardized precipitation index (SPI)[22], the prediction of onset of Australian winter rainfall by RF[23], application of RF to daily and monthly rainfall forecasting[24], hourly rainfall forecasting by SVR[25], reservoir inflow forecasting[26–28], streamflow/ river stage forecasting[29,30], typhoon flood forecasting[31], and hydrologic time series analysis[32]. Hence, the successful application of these two ML algorithms for constructing prediction models in different fields of studies encourages the idea to use the same for long-lead prediction of the EM events.

In a nutshell, the importance of long-lead prediction of EM events to understand the resulting climatic impacts and teleconnection patterns and the lack of prediction skill of existing climatic operational systems, build the motivation for the paper. Based on the motivation, the objective of the paper is to evaluate the ability of ML algorithms to provide effective long-lead prediction of EM events, the first of its kinds, using slowly varying climatic variables as predictors. The following sections provide a detailed description of the results obtained, data used and mathematical description of the models in the methodology.

Results

Identification of the input regions. At the outset, the slowly varying climatic variable, global monthly SSTA, sea surface height anomaly (SSHA) and soil moisture content anomaly (SMC) at 100–289 cm depth, for the period of 1982 to 2017, are selected as the predictors for EMI. Lagged correlations between the monthly observed EMI and SSTA, SSHA and SMC are determined by Kendall's tau (τ) considering the lags of 6, 12, 18 and 24 months to identify the regions significantly (at 1% significance level) associated with the EMI. The statistical significance is computed at 1% significance level after a field significance analysis using two-tailed Z-test. Distribution of Kendall's tau approximately follows normal distribution for large number of sample. The Z-test was performed to find out the statistical significance since the data size was sufficiently large ($>$400). The consideration of (τ) helps to deal with the non-linear relationship between the input and the target. The domain of interest is selected based on the (τ) values, which are statistically significant at 1% significance level. The identification of the significant zones indicates that there can be several numbers of identified input zones to characterize the EMI. However, multi-dimensionality can hinder the accurate interpretation of the effective information and thus dimensionality reduction is always helpful in the prediction process. We used the Supervised Principal Component Analysis (SPCA), which is one of the most effective tools for dimensionality reduction[33,34]. The SPCA utilizes the Hilbert–Schmidt Independence Criterion (HSIC) and develops the principal components based on an orthogonal transformation of the input matrix[35]. By applying SPCA on the n-dimensional input set, a set of principal components is obtained in the order of its association with the target variable. Thus, the first component is expected to exhibit maximum association with the target variable. The mathematical descriptions of Kendall's tau and SPCA technique are provided later in the methodology section. In the present study, the optimal number of principal components to be considered is determined by examining the variation of the prediction performance with number of principal components considered before feeding to the ML tools.

Figures 1 and 2 show the identified zones of SSTA, SSHA and SMC values based on the correlation analysis and the SPCA coefficient values (the associated bar plots) of each identified regions for the leads of 6 and 12 months respectively. The same for the leads of 18 and 24 months can be found in Figs. S1 and S2 in the supplementary document. The latitude longitude information along with their monthly variance values for all the four lead times are represented in Tables S1–S4 respectively in the supplementary document. The squared values of the SPCA coefficients quantify the individual contribution of each selected zones to predict EMI for that particular lead time. A comparison among the different variables based on the values of the SPCA coefficients indicates that the SSTA fields provide maximum information to predict the EMI for all leads. The SSTA fields at Central and North Pacific region show the maximum positive correlations with the target, which is also supported by the highest SPCA coefficient values for lead 6. The contribution from the Central Pacific region decreases at higher lead time i.e. at 12, however, it still shows the maximum association with the target compared to other selected predictors. Nevertheless, the SSTA field from the same region does not show any association at all with target at the leads of 18 and 24 months. ENSO Modoki evolution in tropical Pacific basically determines the EMI relationship with SSTA. The signals from the western Pacific move to central Pacific in 6 to 12 months (e.g. Ashok *et al.* 2007). Therefore, the association in terms of correlation with EMI are not seen in central tropical Pacific at 18 to 24 months lead. On the other hand, the SSTA field over the North Pacific region is significantly correlated to EMI even at higher leads i.e. for the leads of 12, 18 and 24 months although the association is visibly decreasing (in terms of correlation) with the increase in leads. The SSTA over the North Atlantic shows significant negative correlation with the EMI at all the leads. The evolution of ENSO Modoki is associated with several tropical and extra-tropical processes (including the signals on the trails of previous ENSO and ENSO Modoki events). Some of the signals seen in the North Pacific are related to these processes. However, this study lacks the scope to discuss ENSO Modoki evolution mechanism as the focus here is only to evaluate ML approaches for long-lead prediction of EMI. The remaining identified SSTA regions are comparatively lesser contributing, however, still higher than other two climatic variables i.e. SSHA and SMC.

SSHA is found to be the second most contributing climate variable although the contribution is much lesser than SSTA. The SPCA coefficient values range from 0.003 to 0.2 for all the leads considering all the identified fields. At the lead of 6 months, the highest positive correlation of SSHA field at Central Pacific region with the EMI is explicit. Contrastingly, the SSHA field from the Western Pacific region is negatively associated with the EMI, though it has the highest contribution to the prediction among all the identified SSHA zones for lead 6. At the next lead i.e. at lead 12, the Central Pacific SSHA field shows maximum association. All the considered

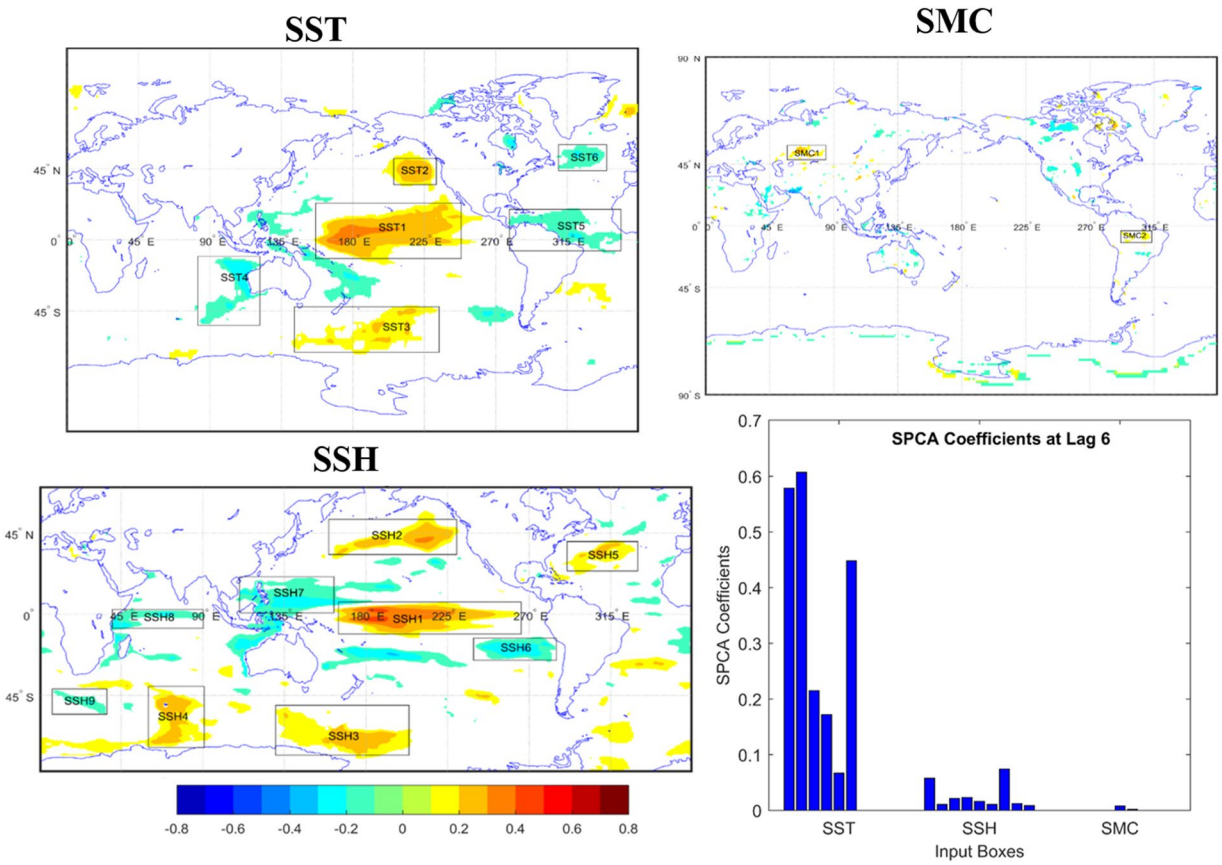

Figure 1. Identified significant zones from the global fields of SSTA, SSHA and SMC along with the SPCA coefficients of each identified zones at 6 months lead.

SSHA fields show trivial associations to predict EMI at lead of 18 months, which remain same for the lead of 24 months except the fact that the association from the Central Pacific region significantly increases perhaps owing to the resultant effect of a previous El Niño Southern Oscillation (ENSO) or EM event. Finally, considering the SMC fields, the contribution of the two identified zones in Europe and Amazon region are visibly insignificant. However, the positive correlation value (Kendall's tau ~ 0.3–0.4) and SPCA coefficient values for the SMC field at European region remain almost same for all the leads.

As stated earlier, the study aims to minimize the multi-dimensionality problem by using the SPCA technique. It quantifies the individual contribution of each selected predictors for EMI. However, the final selection of the input fields is based on the variation of prediction performance with number of input variables considered. It has been observed that for all the leads, the prediction performances increase with the increase of number of input fields. Hence, for the present study, all the identified predictor fields except the SMC over the Amazon region, are used to develop the prediction model. The inclusion of SMC over the Amazon region does not improve the model performances at any lags and thus discarded from the set of input variables. Although, the study lacks the ability to provide any physical justification for this currently, it can be considered as a future study, as the current one emphasises on exploring the EMI predictability using ML algorithms.

Model performance. The performances of the selected models are evaluated by different performance statistics namely Correlation Coefficient (CC), Refined Degree of Agreement (D_r), Root Mean Square Error (RMSE) and the unbiased Root Mean Square Error (uRMSE). The models have been developed independently for all the leads i.e. 6, 12, 18 and 24 months. The independently developed models for each lead are applied and the outcomes are evaluated through performance statistics evaluated during both the development and testing periods.

We applied a 5-month Moving Average (MA) to the EMI to filter out the high frequency intra-seasonal variations. Figures 3 and 4 show the comparison of model performances for the SVR and RF at the lead of 6 and 12 months during model development and testing periods with and without the application of the low-pass filter (Tables S5 and S6 show the values of the model performance metrics for the two cases in the supplementary document). The application of the low-pass filter enhances the skill scores of the ML models. However, the two cases are found to exhibit similar skills in performance metrics across the folds and leads. Hence, most of the conclusions discussed in the following section hold true for both the cases except for some of the instances that are discussed especially wherever necessary. The outstanding model performances for both the SVR and RF during model development and testing periods at lead 6 are explicitly visible from the metrics values for both the cases introduced above. At the lead 12 months, the performance of RF decreases drastically during the testing period,

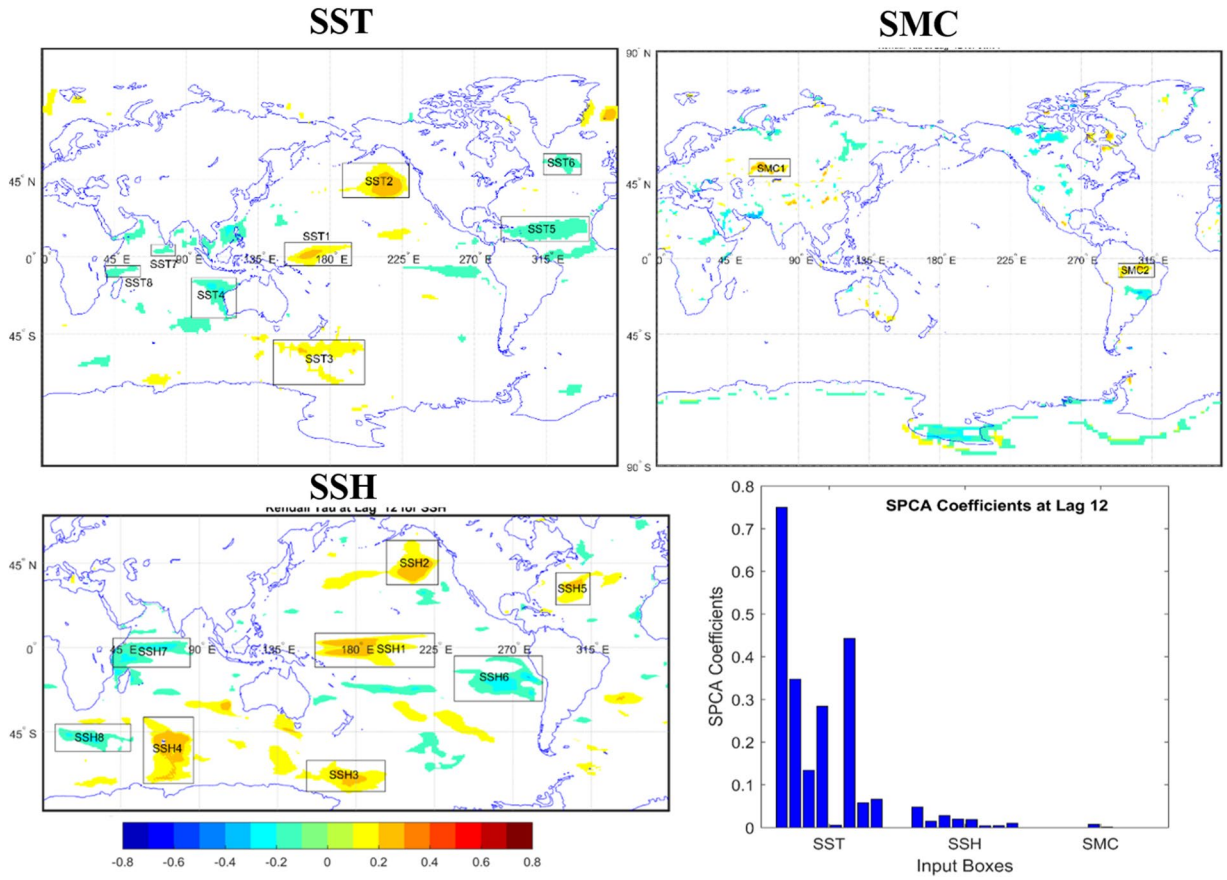

Figure 2. Same as Figure 1 but at 12 months lead.

although the model development shows an acceptable result. However, in case of SVR, although the model performance decreases both for development and testing periods at the lead of 12 months, the performances are reasonably acceptable considering such a sufficiently long lead time. The next lead i.e. 18 months, shows the worst performances for both the SVR and RF perhaps owing to the overtraining of both the models. Nevertheless, the performances for the two models improve in terms of overfitting for the lead of 24 months though the model performances are not acceptable. The performances for the two cases of with and without the short-term fluctuations for the leads of 18 and 24 months can be found in Figs. S3 and S4 in the supplementary document. Briefly, the comparison among different lead times shows that the SVR is able to skilfully predict the EMI up to a lead of 12 months. In case of RF, although the model is able to predict quite accurately at the lead of 6 months, its performance is not good at the lead of 12 months. Beyond that, i.e. for the leads of 18 and 24 months, both the models loose skills to predict owing to too long lead time to capture the variation.

The overall performances of the models at all the leads are further evaluated by computing the model performance metrics considering the data from the development and testing periods of each fold as one time series. Also, the correlation coefficient value between EMI of testing period at lead 6 and that at lead 12 is computed as 0.787 (shown as average of all four folds). Figure 5 shows the overall performances for those two cases with and without the short-term fluctuations for the leads of 6 and 12 months. The same for the leads of 18 and 24 months are provided in the supplementary document in Figure S5. The same inference as discussed above is drawn from both the figures; i.e. the model performances are acceptable up to the lead time of 12 months. Additionally, the model performance is compared with the persistent of EMI, which is computed as follows: the observed EMI is used as the predicted values at the lead of *N*-months (6 or 12 months as examples) and correlated with the actual observed values of that month for the evaluation. Following this, the correlation coefficients for the leads of 6 and 12 months are found to be 0.596 and 0.271 respectively. Whereas, the correlations obtained from SVR and RF predicted EMI range between 0.735 to 0.952 and 0.683 to 0.990 respectively for the lead of 6 months in development and testing periods. The same metrics, for the lead of 12 months are 0.626 to 0.946 and 0.476 to 0.927 for SVR and RF respectively. Evidently, the correlations obtained from model predicted EMI are much higher than the persistence of correlation of EMI for the lead time of 6 and 12 months.

The comparison among the two models indicates the better performance of SVR for all the leads compared to RF. In RF, the peaks are captured very well in model development period. However, the ability to capture the amplitude of the extreme variabilities reduces in testing periods for all the folds and leads, although still able to capture the phases. The difference between the model performance during development and training periods can be reduced to some extent by tuning the m_{try} number and *nodesize* for each case. However, the tuning process was investigated and found to produce the predicted values almost equal to the mean of the observed values without

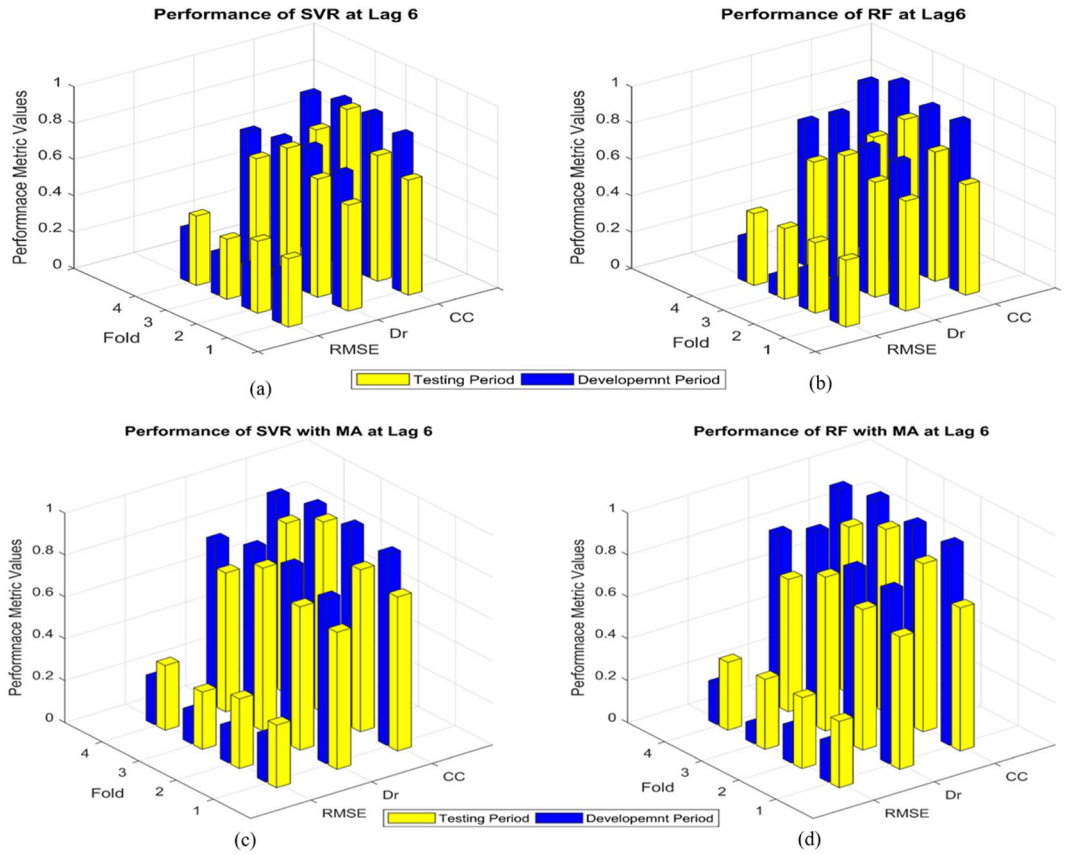

Figure 3. Comparison of model performances at 6 months lead time: (**a**) Performance of SVR without applying the MA; (**b**) Performance of RF without applying the MA; (**c**) Performance of SVR after applying the MA; and (**d**) Performance of RF after applying the MA.

capturing the variations and the extreme values. This could perhaps be due to the short length of the observed data. It may be noted that the model development criteria are maintained uniformly for both RF and SVR to capture the variation and extreme values of EMI at an acceptable level. In case of SVR, although the performance metric values are lesser than RF during the development period, there is a parity of performance between development and testing periods and it is able to capture the peaks reasonably well in the testing periods also. Considering all these issues, SVR is found to be a better option for the prediction of EMI even at higher leads. The application of these two approaches in different fields of studies and comparisons of their performances show that the advantages of these two models, while compared to each other, are very much problem dependent[36]. In our study, the comparative analysis of the performances of the two models indicates the suitability of SVR for predicting EMI at 6 and 12 months lead with short data period.

Figures 6 and 7 show the time series plots of observed and predicted EMI for all the models and folds for the leads of 6 and 12 months for the case of without short-term fluctuations. The counterparts i.e. the time series plots of observed and predicted EMI for all the models and folds for the leads of 6 and 12 months with short-term fluctuations are shown in Figures S6 and S7 in the supplementary document. The comparison of these two cases shows that although the model performances are improved after reducing the short-term fluctuations, the models' abilities in terms of capturing the peaks are not improved significantly. The model performances are not uniform across the four folds. It is observed that for the first fold, where the model development period is approximately 1982 to 2009 and testing period is from 2009 to 2017, the SVR and RF perform the best, especially in testing periods. Particularly, the SVR is able to capture the EM events for both the development period (approximately the years of 1983 and 1999) and testing period (the year of 2010). The SVR-predicted EMI also shows a good association with the observed data for the frequent and shorter peaks throughout the length of the time series. On the other hand, the RF shows an overfitting tendency, i.e. it captures the Modoki events well during the model development period but fails to do so during testing period where it only captures the direction/phase of the peak. The second fold, where the EMI data from the year 1991 to 2017 (approximately) is used to develop the models, shows a similar trend to that of fold 1, although the time series plots show a little deterioration in the ability to capture the shorter as well as extreme peaks for both the SVR and RF models. The third and the fourth folds, give EMI prediction without capturing any of the peaks. At the lead of 12 months, the abilities to capture the peaks decrease drastically for both the models. However, the SVR shows a superior skill than RF to capture the higher peaks of EMI in the testing periods, particularly for the first and the second folds. Similar to the lead of 6 months, the third and fourth folds show not so good model performances for both SVR and RF at leads of 12 months. It is interesting to note that both SVR and RF models could capture the correct phases of the EMI events in all the

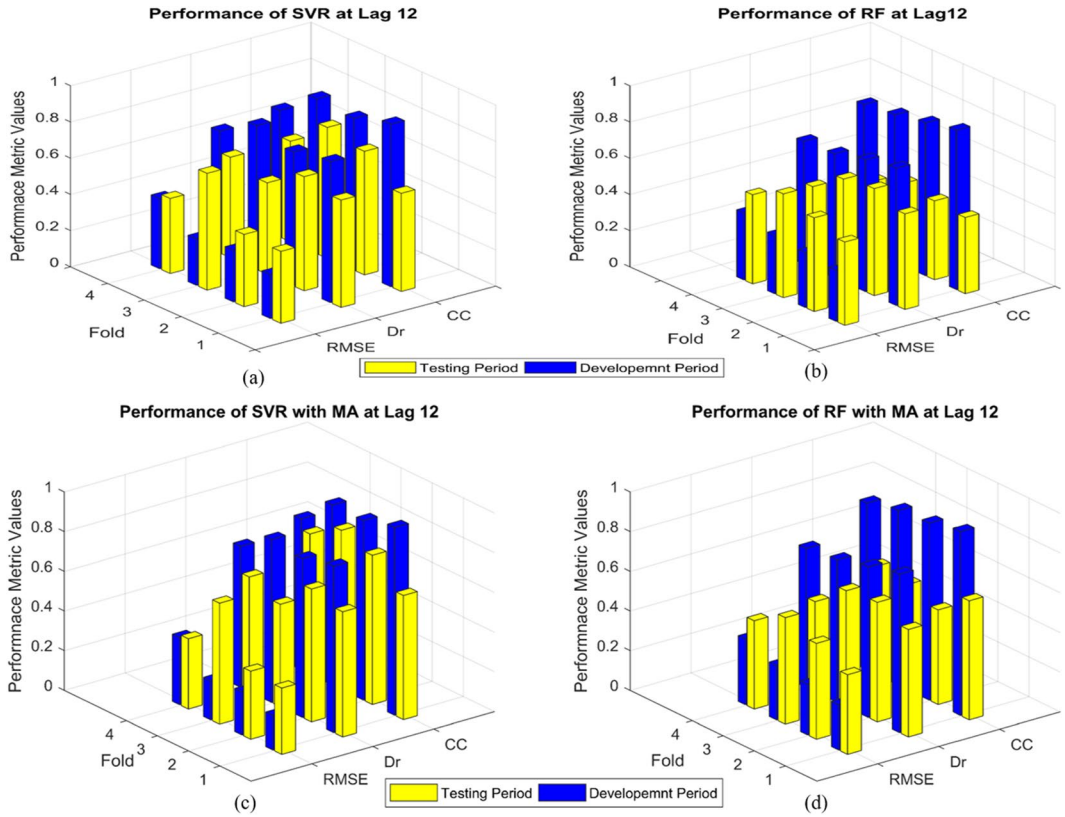

Figure 4. Comparison of model performances at 12 months lead time: (**a**) Performance of SVR without applying the MA; (**b**) Performance of RF without applying the MA; (**c**) Performance of SVR after applying the MA; and (**d**) Performance of RF after applying the MA.

four phases in most of the years. However, the amplitudes of the EMI are underestimated by both the models. The above discussion of different trends of model performances across the four folds are same for the cases of with and without short term fluctuations as mentioned earlier in the paper. The prediction at 18 and 24 months leads are not acceptable, yet for both the leads the model performances follow the same trend of performing better in first and second folds. This can be perceived in the comparison of observed and predicted EMI for these two leads in Figures S8–S11, provided in the supplementary document.

Next, observed Modoki events away from its mean by more than one standard deviation (STD) are specifically picked out for case by case evaluation of prediction performance by SVR. Table 1 shows such periods, observed phases of the Modoki events (positive/negative) along with the predicted phases in case of 6-month as well as 12-month lead times. The prediction ability is evaluated in following four categories:- (a) both phase and magnitude are accurately (away from mean by more than one STD) predicted and designated as ($\surd\surd\surd$); (b) phase is correctly predicted but the magnitude is predicted marginally lesser than one STD which is designated as ($\surd\surd$); (c) phase is correctly predicted but the magnitude is predicted poorly, i.e., much less than the STD value ($\surd$); and (d) incorrect prediction of phases and amplitudes ($\times$). For the lead of 6 months, more accurate predictions of magnitudes and phases are noticed in 1983–1984, 1991–1992, 1998–1999, and 2000. In the years of 1988–1989, 1990–1991, 1994–1995, 2001, 2008–2009, and 2011–2012, predictions of correct phases are noticed but relatively poorer performances in terms of marginally underestimated magnitudes, i.e. falling in the second category of the prediction ability. The phases are predicted correctly but poor predictions of magnitudes are observed for the events occurred in the years 1998, 2006, 2009–2010, 2015 and 2016. For the lead time of 12 months, in general, the overall performance is poorer than 6-month lead time prediction as expected. However, still the ability to capture both phases and magnitude in some of the cases (1991–1992, 1994–1995 and 2011–2012) is noted. The prediction ability falls in the second category in the years of 1983–1984, 1988–1989, 1990–1991, 2008–2009, 2010–2011 and 2015 when the predicted magnitudes marginally fail to capture one STD away values but captures the exact phases of the Modoki events. In a nutshell, for both the lead times, 45% of the prediction cases fall in the first or second category i.e. almost correctly predict both the amplitudes and phases. The rest of the identified time periods fall either in third ($\surd$) or fourth ($\times$) category of prediction, which may be understandable, considering the long lead times of 12 months.

Thus, the advantage of prediction using ML approach to determine the correct phases of the Modoki events with long lead-times (6 and 12 months) is creditable as compared to the existing climate forecast models as discussed before in Section 1. Moreover, the predictions of amplitudes, specifically from SVR, are notably advantageous in most of the cases with 6-month lead time, whereas the existing prediction models are strictly limited to distinguish the patterns 0–2 months ahead.

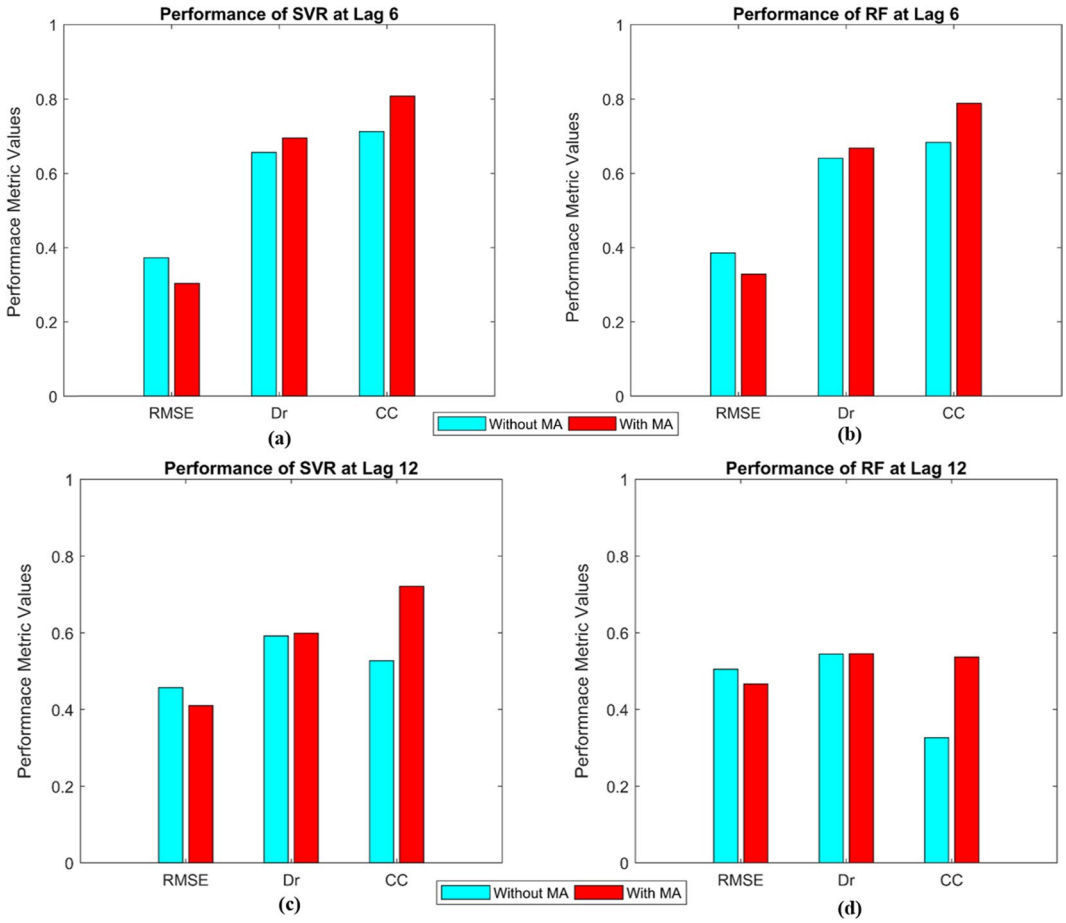

Figure 5. Comparison of overall model performances without and with applying the MA to remove short-term fluctuations: (**a**) Performance of SVR at lead time of 6 months; (**b**) Performance of RF at lead time of 6 months; (**c**) Performance of SVR with lead time of 12 months; and (**d**) Performance of RF with lead time of 12 months.

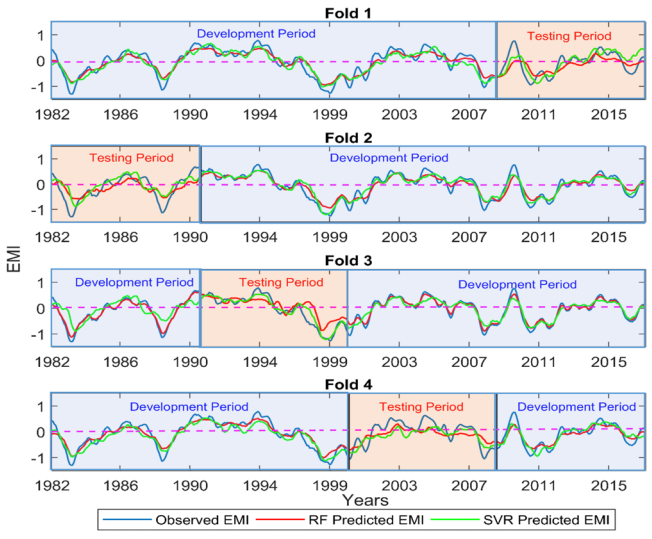

Figure 6. Comparison of observed EMI with the RF-predicted and SVR-predicted EMI at 6 months lead time after removing short-term fluctuations (5-month moving average).

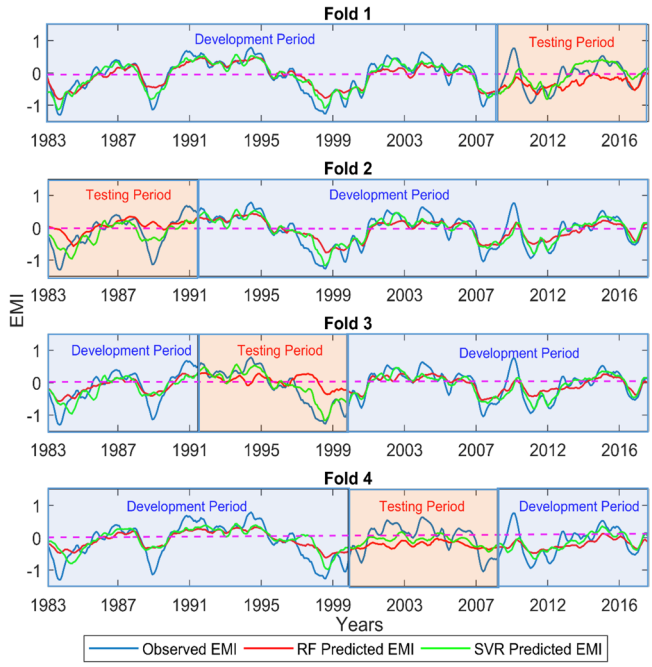

Figure 7. Comparison of observed EMI with the RF-predicted and SVR-predicted EMI at 12 months lead time after removing the short-term fluctuations (5-month moving average).

Periods	Phase	Prediction Lead Time 6-month	12-month
April, 1983–April, 1984	Negative	√√√	√√
October, 1988–June, 1989	Negative	√√	√√
November, 1990–June, 1991	Positive	√√	√√
October, 1991–January, 1992	Positive	√√√	√√√
August, 1994–June, 1995	Positive	√√	√√√
October, 1997–December, 1997	Negative	×	×
January, 1998–May, 1998	Negative	√	×
June, 1998–September, 1999	Negative	√√√	√
January, 2000–June, 2000	Negative	√√√	√
January, 2001–May, 2001	Negative	√√	√
August, 2002–October, 2002	Positive	×	×
August, 2004–December, 2004	Positive	×	×
March, 2006–April, 2006	Negative	√	×
January, 2008–March, 2009	Negative	√√	√√
November, 2009–March, 2010	Positive	√	×
September, 2010–June, 2011	Negative	√	√√
December, 2011–June, 2012	Negative	√	√√√
March, 2015–April, 2015	Positive	√	√√
November 2016	Negative	√	√
March, 2017–April, 2017	Negative	×	√

Table 1. Case by case evaluation of prediction performance by SVR. The table shows the periods of Modoki events away from its mean by more than one standard deviation and corresponding observed and predicted phases (positive/negative). The prediction ability is evaluated in four categories i.e.: (a) both phase and magnitude are accurately (away from mean by more than one STD) predicted, designated as (√√√); (b) phase is correctly predicted but the magnitude is predicted marginally lesser than one STD, designated as (√√); (c) phase is correctly predicted but the magnitude is predicted poorly, i.e., much less than the STD value (√); and (d) incorrect prediction of phase and amplitude (×).

Conclusions

Long-lead (6 to 12 months) prediction of EMI with reasonable accuracy is achieved in this study using two ML algorithms namely SVR and RF. The input to the models were the anomalies of the slowly varying climatic

variables such as SSTA, SSHA and SMC. Firstly, the correlation analysis with Kendall's tau helps to identify the significantly contributing signals from global anomaly fields of each predictor considering the non-linear dynamics. Subsequently, the study uses the SPCA technique to reduce the dimensionality which also ensures the selection of predictors having maximum association with the target i.e. the EMI. The SPCA analysis shows the coefficients corresponding to SSTA fields have the highest contributions in EMI predictions as compared to that of SSHA and SMC. SSTA from the Central and Northern Pacific regions along with the signal from Northern Atlantic region are having the maximum association with the EMI. While identifying, along with the already established fields connected with the EMI such as SSTA and SSHA from central pacific region, some additional fields are identified. These additional fields such as, SSTA signals from Northern Pacific region and Northern Atlantic region, SSHA signal form Indian Ocean region and SMC signal from Europe, are found to have significant correlation even at the higher leads of 12, 18 and 24 months. However, enhancement of model performances allows the study to include almost all the SSTA, SSHA and SMC fields despite having low SPCA coefficients. The predictor selection leads to the development of SVR and RF prediction models at the leads of 6, 12, 18 and 24 months.

The results reveal that the SVR gives consistently better performance for all the leads as RF exhibits a tendency to overfitting. Regarding the leads, both the SVR and RF show excellent performances for the 6-month ahead prediction. The predictability decreases for 12 months lead prediction for both the models though the phases of Modoki events are still captured properly. Yet, the model performance of SVR is reasonably well considering the leads whereas the RF consistently shows the problem of overfitting. However, the performance of RF has improved at the lead of 12 months by reducing the short-term fluctuations. The skills of 18 and 24 months lead predictions are not acceptable and thus the idea of developing prediction models at those leads is discarded. Overall, the study concludes that a skillful long-lead prediction, i.e. 6 to 12 months ahead prediction, of El Niño Modoki events is possible with the ML algorithms such as SVR using the SSTA, SSHA and SMC fields as the predictors.

Data. The following data sets for the period 1982–2017, are used in this study: a) monthly Sea Surface Temperature Anomaly (SSTA) data from Optimum Interpolation Sea Surface Temperature (OISST) from National Centers for Environmental Information (NOAA) at a spatial resolution of $1.0° \times 1.0°$; (b) monthly Sea Surface Height Anomaly (SSHA) data from NCEP Global Ocean Data Assimilation System (GODAS) at a spatial resolution of $0.5° \times 0.5°$ and its accuracy is established by several previous studies[37–39]; and c) monthly Soil Moisture Anomaly (SMC) data at 100–289 cm depth from ERA Interim Data from European Centre for Medium Range Weather Forecast (ECMWF) at a spatial resolution of $0.75° \times 0.75°$. The accuracy and quality of the SMC product has been investigated in several studies based on the ground observations in recent past for surface as well as root zone depths[40–42].

The target variable i.e. the EMI is available and thus obtained from the website of Japan Agency for Marine-Earth Science and Technology (JAMSTEC; http://www.jamstec.go.jp/aplinfo/sintexf/DATA/emi.monthly.txt). The EMI can be represented by the following equation,

$$\mathrm{EMI} = [\mathrm{SSTA}]\mathrm{A} - 0.5 \times [\mathrm{SSTA}]\mathrm{B} - 0.5 \times [\mathrm{SSTA}]\mathrm{C} \quad (1)$$

The Eq. (1) represents the area-averaged SSTA over each of the region A(bounded by 165°E–140°W, 10°S–10°N), B (bounded by 110°W–70°W, 15°S–5°N), and C (bounded by 125°E–145°E, 10°S–20°N), respectively.

Methodology

The overall methodology consists of mainly two steps. The first step is to identify and select the regions for each input variable i.e. the SSTA, SSHA and SMC which are highly associated with the target variable EMI. The initial selection of associated input regions is based on lagged correlation analysis using Kendall's tau at the specified leads of 6, 12, 18 and 24 months at the significance level of 0.01. The areas of significant correlations are investigated for each fold and each lag individually (however, not shown to avoid redundancy), and the most common areas identified for all the folds are selected as the final contributing areas for the predictors. Subsequently, the study attempts to deal with multi-dimensionality problem using the Supervised Principal Component Analysis (SPCA). Although, as discussed above, the final selection of the contributing input zones is based on the model performances. After that, the final step of the study is the prediction model development considering the selected input variable zones and comparison of the model performances at the leads of 6, 12, 18 and 24 months using two ML approaches i.e. the Support Vector Regression (SVR) and Random Forest (RF). The mathematical descriptions of all the steps are elaborated in the following sections.

Predictor selection based on Kendall's Tau and Supervised Principal Component Analysis. The associated zones of the SSTA, SSHA and SMC are identified using the Kendall's Tau (τ) as discussed above. It is a rank-based, non-parametric statistical measure which is defined by the difference between the probability of concordance and discordance of two random variables[43]. Suppose, V and Y are the input variable and EMI respectively. Mathematically, Kendall's Tau can be represented by following equation:

$$\tau = P[(V_i - V_j)\,(Y_i - Y_j) > 0] - P[(V_i - V_j)\,(Y_i - Y_j) < 0] \quad (2)$$

where, i and j are any two time steps which are not equal ($i.e.\ i \neq j$).

After identifying the statistically significant associated zones at 1% significance level, the study intends to diminish the redundancy of information due to multi-dimensionality using Supervised Principal Component Analysis (SPCA) performed on the development period dataset. Let, a set of n observed data points during development period each comprising of p characteristics form a matrix, X of $p \times n$ dimension and Y is the $1 \times n$

dimensional matrix of the output variable. The SPCA technique aims to find the subspace U^TX to maximize the association between the output variable Y and the projected input matrix $U^T X$ using HSIC, where U is an orthogonal projection matrix of size $p \times 1$. The orthogonal transformation matrix, U which maps the data points to a space where features are not correlated, is solved by the following optimizing problem,

$$\underset{U}{\arg\max}\, tr(U^T XHLHX^T U),\ subject\ \ to\colon UU^T = 1 \tag{3}$$

where, the $\underset{U}{\arg\max}$ indicates a maximization problem considering U as an argument. The symmetric and real matrix $Q = XHLHX^T$ of size $P \times P$, has P number of eigenvalues ($\lambda_1 \leq \ldots \leq \lambda_p$) and corresponding eigenvectors [$\nu_1, \ldots, \nu_p$], each consisting of P number of elements. Generally, maximum value of the cost function is $\lambda_p + \lambda_{p-1} + \ldots + \lambda_{p-d+1}$ and the optimum solution is $U = [\nu_p, \nu_{p-1}, \ldots, \nu_{p-d+1}]$, where d is the dimension of $[U^TX]$. Hence, $U = [\nu_p]$, which produces the coefficients for P different input variables and ensures the best association of the product to the output variable. Physically, the coefficients provide information on the weightages for each of the considered input variables. The square of the SPCA coefficients represent the contribution of each input variable to estimate the target output and the sum of the squares of the SPCA coefficients is equal to one. Therefore, the comparison of the absolute values of the SPCA coefficients corresponding to each of the specific input helps to select the best possible combination of inputs for EMI prediction. Additionally, it ensures the selected combination of the input variables have the maximum association with the target variable.

Support Vector Regression (SVR). Support Vector Regression (SVR) has been popular in many disciplines nowadays which uses a penalty term added to the error function to penalize the resultant complexity. It aims to decrease dthe problem of overfitting by adopting the theory of structural risk minimization. The current study uses the SVR for constructing the EMI prediction models at the leads of i.e. 6, 12, 18 and 24 months. A brief mathematical description of SVR is as follows.

Let, [$(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots (x_l, y_l)$] be a training dataset where x_i is an input vector with its corresponding output vector y_i and l is the number of data pairs. The SVR finds a regression function $f(x) = \langle w, x \rangle + b$ to represent the dependency that best describes the observed output y with an error tolerance ε, where w and b are the weighting vector and bias respectively. For this purpose, the original input domain is mapped onto a higher dimensionality space, where the function underlying the data is assumed to be linear. The SVR problem in the transformed space is identified by solving the following optimization problem,

$$\begin{aligned} &Minimize \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{L}(\xi_i + \xi_i^*) \\ &Subject\ \ to \quad \begin{cases} Y_i - \sum_{j=1}^{K}\sum_{i=1}^{L} w_j x_{ji} - b \leq \varepsilon + \xi_i, \\ \sum_{j=1}^{K}\sum_{i=1}^{L} w_j x_{ji} - y_i \leq \varepsilon + \xi_i^*, \\ \xi_i, \xi_i^* \geq 0, \end{cases} \end{aligned} \tag{4}$$

where, ε is the Vapniks insensitive loss function when data are outside of the tube of error tolerance; C is the capacity parameter cost which is a positive constant that determines the degree of penalized loss when a training error occurs to tune the trade-off between model complexity and tolerance to empirical errors; and ξ_i and ξ_i^* are called the slack variables which measure the distance (in the target space) of the training samples lying outside the ε-insensitive tube from the tube itself[44]. The functional dependency *f(x)* can be written as,

$$f(x) = \sum_{j-1}^{K} w_j x_j + b \tag{5}$$

where, K is the number of support vectors.

The optimization problem is solved using the dual formulation subject to constraints in the loss function and introducing the Lagrange multipliers, α_i and α_i^*. By solving the optimization problem the final prediction function is:

$$f'(x) = \sum_{i \in N}(\alpha_i - \alpha_i^*) k(x_i, x) + b \tag{6}$$

where, $k(\ldots)$ is kernel function which computes non-linear dependence between the two input variables x_i and x where x_i are the "support vectors" and b is the bias. In the present study, the Radial Basis Function (RBF) kernel is used in the prediction of EMI. It is proven the best among several possibilities for the choice of kernel function, including linear, polynomial, sigmoid and splines, because of its excellent performance in capturing nonlinear relationship[45,46]. Mathematically, the RBF with kernel width $-\gamma$, can be represented as,

$$k(x_i, x) = \exp(-\gamma\|x - x_i\|^2, \qquad \gamma > 0 \tag{7}$$

Parameters	Lead 6				Lead 12				Lead 18				Lead 24			
	Fold				Fold				Fold				Fold			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Support Vector Regression Parameters																
Cost (C)	50	5	20	20	10	18	5	5	1	4	4	4	10	10	25	10
Kernel Width (γ)	0.01	0.01	0.01	0.005	0.01	0.005	0.009	0.001	0.05	0.009	0.002	0.001	0.001	0.001	0.1	0.01
Insensitive parameter (ε)	0.9	0.1	0.5	0.9	0.4	0.4	0.01	0.009	0.01	0.1	0.01	0.09	0.001	0.09	0.1	0.09
Random Forest Parameters																
n_{tree}	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500
m_{try}	5	13	10	13	13	10	19	3	1	10	5	3	10	10	5	10
nodesize	40	40	3	50	56	66	98	100	10	20	50	40	60	40	50	80

Table 2. Tuning parameters of SVR and RF Model for different folds and different leads.

Random forest (RF). RF is one of another ML algorithm for predictive analytics consisting of an ensemble of simple trees. The two major components of RF algorithm are: (1) randomness and (2) ensemble learning.

1. *Randomness*

- n_{tree} bootstrap samples are randomly selected from the data set of size N with M features with replacement. For each bootstrap, approximately two thirds of the entire dataset is chosen as a subset (i.e. around one-third of the subset are replicated in the subset) to develop the decision tree model. The un-chosen one-third samples in the original dataset are called out-of-bag (OOB) data. This OOB data is used to get unbiased estimates of the regression error and the importance of the variables used for constructing the tree.
- For each of the bootstrap samples a regression tree is grown as such that at each node, a subset of the predictor variables ($m_{try} < M$) is randomly selected to generate the binary rule to make the decision for the best split. The predictor with the lowest residual sum of squares is selected for the split. Tuning of this parameter is needed for optimal performance although it is not very sensitive to the model performance.

2. *Ensemble learning*

- A subset of size N' (bootstrap sample) with m_{try} features is drawn after random selection process to construct a single decision tree to the largest extent possible without pruning for each of the n_{tree} tree.
- Finally, predictions are calculated as all the n_{tree} trees vote upon the observation of the test data set or the OOB observation. In this ensemble learning method each of the decision trees inside the ensemble contributes individually. The final estimate is obtained by averaging the results from individual trees[20].

Model tuning. For all ML algorithms, *k*-fold cross-validation method is performed to determine the optimal model settings and evaluating the generalized model performance to an independent data set[47,48]. The *k*-fold cross validation also helps to avoid overfitting. To apply the *k*-fold cross validation, the dataset is randomly partitioned into 4 equally sized folds for the present study. Thus, every fold is a subset (1/4) of the complete time series. Models were then fitted by repeatedly leaving out one of the folds. The models are tuned individually for all the folds for all the four leads. A model's performance is determined by predicting on the fold left out.

The RF implementation of the "randomForest" package[49] in R was applied. The number of predictor variables randomly selected at each split (m_{try}) was tuned for each value between one and the number of input variables[19]. The number of trees (n_{tree}) was set to 500 after no increase of accuracy was observed after 500 trees. The "e1071" package[50] in R provided the SVR algorithm used in this study. The cost, gamma and ε-insensitive loss function values were tuned for 2 to 512; 0.001 to 1 and 0.001 to 1 respectively. A radial kernel function was used to account for non-linearity. Table 2 shows the values of SVR and RF parameters for all the folds and leads considered in the study.

Received: 29 May 2019; Accepted: 16 December 2019;
Published online: 15 January 2020

References

1. Ashok, K., Behera, S. K., Rao, S. A., Weng, H. & Yamagata, T. El Nino odoki and its possible teleconnection. *J. Geophys. Res.* **112**, C11007, https://doi.org/10.1029/2006JC003798 (2007).
2. Weng, H., Behera, S. K. & Yamagata, T. Anomalous winter climate conditions in the Pacific rim during recent El Nin˜o Modoki and El Nin˜o events. *Clim Dyn* **32**, 663–674, https://doi.org/10.1007/s00382-008-0394-6 (2009).
3. Cai, W. & Cowan, T. La Nin˜a Modoki impacts Australia autumn rainfall variability. *Geophys Res Lett* **36**, L12805, https://doi.org/10.1029/2009GL037885 (2009).
4. Taschetto, A. S. & England, M. H. El Nin˜o Modoki impacts on Australian rainfall. *J Clim* **22**, 3167–3174, https://doi.org/10.1175/2008JCLI2589.1 (2009).
5. Ratnam, J. V., Behera, S. K., Masumoto, Y., Takahashi, K. & Yamagata, T. Anomalous climatic conditions associated with the El Niño Modoki during boreal winter of 2009. *Clim Dyn* **39**(1–2), 227–238 (2011).

6. Ratnam, J. V., Behera, S. K., Masumoto, Y. & Yamagata, T. Remote effects of El Niño and Modoki events on the austral summer precipitation of Southern Africa. *J Clim* **27**, 3802–3815 (2014).
7. Sahu, N. *et al*. El Niño Modoki connection to extremely low streamflow of the Paranaíba River in Brazil. *Clim Dyn* **42**(5–6), 1509–1516 (2014).
8. Behera, S. & Yamagata, T. Climate Dynamics of ENSO Modoki Phenomena. *Oxford Research Encyclopedia of Climate Science*, https://doi.org/10.1093/acrefore/9780190228620.013.612 (2018).
9. Weng, H., Ashok, K., Behera, S. K. & Rao., S. A. Impacts of recent El Nino Modoki on dry/wet conditions in the Pacific ~Rim during boreal summer. *Climate Dyn.* **29**, 113–129, https://doi.org/10.1007/s00382-007-0234-0 (2007).
10. Feng, J., Wang, L., Chen, W., Fong, S. K. & Leong., K. C. Different impacts of two types of Pacific Ocean warming on Southeast Asian rainfall during boreal winter. *J. Geophys. Res.* **115**, D24122, https://doi.org/10.1029/2010JD014761 (2010).
11. Feng, J. & Li, J. Influence of El Nino Modoki on spring ~ rainfall over south China. *J. Geophys. Res.* **116**, D13102, https://doi.org/10.1029/2010JD015160 (2011).
12. Zhang, W., Jin, F.-F., Li, J. & Ren., H. Contrasting impacts of two-type El Nino over the western North Pacific during ~boreal autumn. *J. Meteor. Soc. Japan* **89**, 563–569, https://doi.org/10.2151/jmsj.2011-510 (2011).
13. Zhang, W., Jin, F.-F., Ren, H., Li, J. & Zhao., J. Differences in teleconnection over the North Pacific and rainfall shift over the USA associated with two types of El Nino during boreal autumn. *~J. Meteor. Soc. Japan* **90**, 535–552, https://doi.org/10.2151/jmsj.2012-407 (2012).
14. Yuan, Y. & Yang, S. Impacts of different types of El Nino on East Asian climate: Focus on ENSO cycles. *J. Climate* **25**, 7702–7722, https://doi.org/10.1175/JCLI-D-11-00576.1 (2012).
15. Hendon, H. H., Lim, E., Wang, G., Alves, O. & Hudson., D. Prospects for predicting two flavors of El Nino. *~Geophys. Res.Lett.* **36**, L19713, https://doi.org/10.1029/2009GL040100 (2009).
16. Lim, E. P., Hendon, H. H., Hudson, D., Wang, G. & Alves, O. Dynamical forecast of inter–El Nin˜o variations of tropical SST and Australian spring rainfall. *Mon Weather Rev* **137**, 3796–3810, https://doi.org/10.1175/2009MWR2904.1 (2009).
17. Jeong, H.-I. & Coauthors. Assessment of the APCC couple MME suite in predicting the distinctive climate impacts of two flavors of ENSO during boreal winter. *Climate Dyn.* **39**, 475–493, https://doi.org/10.1007/s00382-012-1359-3 (2012).
18. Sun, Q., Bo, W. U., Zhou, T. J. & Yan, Z. X. ENSO hindcast skill of the IAP-DecPreS near-term climate prediction system: comparison of fullfield and anomaly initialization. *Atmospheric and Oceanic Science Letters* **11**(1), 54–62, https://doi.org/10.1080/16742834.2018.1411753 (2018).
19. Kuhn, M. & Johnson, K. Applied Predictive Modeling. First ed. Springer, New York (2013).
20. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
21. Vapnik, V. N. Statistical Learning Theory. John Wiley & Sons, New York (1998).
22. Chen, J., Li, M. & Wang, W. Statistical uncertainty estimation using random forests and its application to drought forecast. *Math. Prob. Eng.* 915053 (2012).
23. Firth, L., Hazelton, M. L. & Campbell, E. P. Predicting the onset of Australian winter rainfall by nonlinear classification. *J. Clim.* **18**, 772–781 (2003).
24. Taksandel, A. A. & Mohod, P. S. Applications of data mining in weather forecasting using frequent pattern growth algorithm. *Int. J. Sci. Res.* **4**(6), 3048–3051 (2013).
25. Lin, G. F., Chen, G. R., Wu, M. C. & Chou, Y. C. Effective forecasting of hourly typhoon rainfall using support vector machines. *Water Resour. Res.* **45**(8), W08440 (2009a).
26. Lin, G. F., Chen, G. R., Huang, P. Y. & Chou, Y. C. Support vector machine-based models for hourly reservoir inflow forecasting during typhoon-warning periods. *J. Hydrol.* **372**(1–4), 17–29 (2009b).
27. Lin, G. F., Chen, G. R. & Huang, P. Y. Effective typhoon characteristics and their effects on hourly reservoir inflow forecasting. *Adv. Water Resour.* **33**(8), 887–898 (2009b).
28. Nguyen, T. T. An l1-regression random forests method for forecasting of Hoa Binh reservoir's incoming flow. In: Proc. 1st International Workshop on Pattern Recognition for Multimedia Content Analysis, Ho Chi Minh City, Vietnam, 10 October 2015. IEEE Vietnam Section, Vietnam (2015).
29. Chen, S. T. & Yu, P. S. Pruning of support vector networks on flood forecasting. *J. Hydrol.* **347**(1–2), 67–78 (2007).
30. Maity, R., Bhagwat, P. P. & Bhatnagar, A. Potential of support vector regression for prediction of monthly streamflow using endogenous property. *Hydrol. Process.* **24**(7), 917–923 (2010).
31. Lin, G. F., Chou, Y. C. & Wu, M. C. Typhoon flood forecasting using integrated two-stage support vector machine approach. *J. Hydrol.* **486**, 334–342 (2013).
32. Yu, X. Y. & Liong, S. Y. Forecasting of hydrologic time series with ridge regression in feature space. *J. Hydrol.* **332**(3–4), 290–302 (2007).
33. Hong, W. C. & Pai, P. F. Potential assessment of the support vector regression technique in rainfall forecasting. *Water Resour. Manage.* **21**(2), 495–513 (2007).
34. Das, S. K. & Maity, R. A hydrometeorological approach for probabilistic simulation of monthly soil moisture under bare and crop land conditions. *Water Resources Research*, https://doi.org/10.1002/2014WR016043 (2014).
35. Pal, M. *et al*. Satellite based Probabilistic Assessment of Soil Moisture using C-band Quad-polarized RISAT 1 data. *IEEE Transactions on Geoscience and Remote Sensing* **55**(3), 1351–1362, https://doi.org/10.1109/TGRS.2016.2623378 (2017).
36. Barshan, E., Ghodsi, A., Azimifar, Z. & Jahromi, M. Z. Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. *Pattern Recognition* **44**, 1357–1371 (2011).
37. Derber, J. C. & Rosati, A. A global oceanic data assimilation system. *J. Phys. Oceanogr.* **19**, 1333–1347 (1989).
38. Behringer, D. W., Ji, M. & Leetmaa, A. An improved coupled model for ENSO prediction and implications for ocean initialization. Part I: The ocean data assimilation system. *Mon. Wea. Rev.* **126**, 1013–1021 (1998).
39. Behringer, D. W. & Xue, Y. Evaluation of the global ocean data assimilation system at NCEP: The Pacific Ocean. Eighth Symposium on Integrated Observing and Assimilation Systems for Atmosphere, Oceans, and Land Surface, AMS 84th Annual Meeting, Washington State Convention and Trade Center, Seattle, Washington, 11–15 (2004).
40. Jing, W., Song, J. & Zhao, X. Validation of ECMWF multi-layer reanalysis soil moisture based on the OzNet hydrology network. *Water* **10**, 1123, https://doi.org/10.3390/w10091123 (2018).
41. Balsamo, G. *et al*. ERA-Interim/Land: a global land surface reanalysis data set. *Hydrol. Earth Syst. Sci.* **19**, 389–407, https://doi.org/10.5194/hess-19-389-2015 (2015).
42. Albergel, C., de Rosnay, P., Balsamo, G., Isaksen, L. & Muñoz-Sabater, J. Soil Moisture Analyses at ECMWF: Evaluation Using Global Ground-Based *In Situ* Observations. *J. Hydrometeor.* **13**, 1442–1460, https://doi.org/10.1175/JHM-D-11-0107.1 (2012).
43. Verikas, A., Gelzinis, A. & Bacauskiene, M. Mining data with random forests: a survey and results of new tests. *Pattern Recogn.* **44**, 330–349 (2011).
44. Embrechts, P., Lindskog, F., & McNeil, A. Modelling dependence with copulas and applications to risk management, in Handbook of Heavy Tailed Distributions in Finance, pp. 329–384, Elsevier, New York (2003).
45. Ahmad, S., Kalra, A. & Stephen, H. Estimating soil moisture using remote sensing data: A machine learning approach. *Advances in Water Resources* **33**(2010), 69–80 (2010).
46. Ghosh, S. SVM-PGSL coupled approach for statistical downscaling to predict rainfall from GCM output. *J. Geophys. Res.* **115**, D22102 (2010).

47. Shaoa, C. *et al*. Feature Selection for Manufacturing Process Monitoring Using Cross-Validation. Proceedings of NAMRI/SME, 41 (2013).
48. Jiang, P. & Chen, J. Displacement prediction of landslide based on generalized regression neural networks with K-fold cross-validation. *Neurocomputing* **198**, 40–47 (2016).
49. Liaw, A. & Wiener, M. Classification and regression by randomForest. *R. News* **2**, 18–22 (2002).
50. Karatzoglou, A., Wien, T. U., Smola, A., Hornik, K. & Wien, W. kernlab — an S4 package for kernel methods in R. *J. Stat. Softw.* **11**, 1–20 (2004).

Acknowledgements

The research was partly supported by Japan Agency for Marine-Earth Science and Technology (JAMSTEC) under Project-B, an initiative to develop AI techniques for climate predictions. Authors RM and MP are also partly supported by Department of Science and Technology, Climate Change Programme (SPLICE), Government of India (Ref No. DST/CCP/CoE/79/2017(G)) through a sponsored project. Authors are thankful to Prof. Toshio Yamagata for the suggestions to improve the analysis. Authors are thankful to ECMWF for making available the ERA-Interim reanalysis through their web site https://apps.ecmwf.int/datasets/data/interim-full-daily and to NOAA/ESRL PSD, Boulder, Colorado, USA for providing the SST and SSH dataset (http://www.esrl.noaa.gov/psd/data/gridded. The anomalies for the fields were derived using the NCAR Command Language (http://www.ncl.ucar.edu/). All the figures are created using MATLAB. The R statistical software package (https://www.r-project.org) was used in the model development and computations.

Author contributions

SKB conceived the central idea, and RM together with JVR conceptualized the model study. MP and RM carried out the model development, analysis and prepared the manuscript during their visits to JAMSTEC together with JVR. RM, JVR, MN and SKB contributed to the interpretation of results and preparation of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-57183-3.

Correspondence and requests for materials should be addressed to R.M.

Reprints and permissions information is available at www.nature.com/reprints.

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