ENVIRONMENTAL RESEARCH
LETTERS

OPEN ACCESS

RECEIVED
3 May 2021

REVISED
3 June 2021

ACCEPTED FOR PUBLICATION
11 June 2021

PUBLISHED
1 July 2021

LETTER

Seasonal prediction of Indian summer monsoon onset with echo state networks

Takahito Mitsui[1,2,*] and **Niklas Boers**[1,2,3]

[1] Department of Mathematics and Computer Science, Free University of Berlin, Arnimallee 6, Berlin 14195, Germany
[2] Potsdam Institute for Climate Impact Research (PIK), Telegraphenberg, Potsdam 14473, Germany
[3] Department of Mathematics and Global Systems Institute, University of Exeter, Exeter, United Kingdom
[*] Author to whom any correspondence should be addressed.

E-mail: takahito321@gmail.com

Keywords: Indian Summer monsoon onset, seasonal prediction, artificial neural network, echo state network

Supplementary material for this article is available online

Abstract

Although the prediction of the Indian Summer Monsoon (ISM) onset is of crucial importance for water-resource management and agricultural planning on the Indian sub-continent, the long-term predictability—especially at seasonal time scales—is little explored and remains challenging. We propose a method based on artificial neural networks that provides skilful long-term forecasts (beyond 3 months) of the ISM onset, although only trained on short and noisy data. It is shown that the meridional tropospheric temperature gradient in the boreal winter season already contains the signals needed for predicting the ISM onset in the subsequent summer season. Our study demonstrates that machine-learning-based approaches can be simultaneously helpful for both data-driven prediction and enhancing the process understanding of climate phenomena.

1. Introduction

The Indian Summer Monsoon (ISM) typically sets in at the Southern tip of India around the beginning of June. India receives 70–90% of annual rainfall sums during the monsoon season, i.e. during the four months from June to September [1]. The meridional gradient of the tropospheric temperature, ΔTT, over the southern Euracian region (figures 1(a) and S1) is a widely-used thermal index to define the ISM onset [2–5]. The onset (withdrawal) date is defined as the day of year at which ΔTT changes its sign from negative to positive (positive to negative); see figure 1(a). We refer to the Data and Method section for further details on the ΔTT index and comparisons to alternative definitions of ISM onset dates. While the annual monsoon cycle is quasi-regular, the onset dates exhibit interannual variations (figure 1(b)). Based on the ΔTT index, the mean onset date is June 2nd (153rd day in a year) and the standard deviation (s.d.) is 5.8 days over the time interval 1948–2020. However, the range of variation reaches 29 days during this period. An early or delayed onset of the ISM can have severe impacts on rain-fed agriculture because agricultural practices are traditionally tied to the mean onset date [6, 7]. Therefore, accurate prediction of the onset date is of crucial importance for effective agricultural planning, but also more generally for water-resource management on the Indian sub-continent with more than one billion inhabitants.

Empirical statistical modelling and numerical, process-based weather modelling are the two main approaches for the prediction of the ISM onset dates. Statistical models are typically derived via linear regression of onset dates on selected meteorological variables [8–12] or their principal components [13]. The selection of predictors is a crucial part of the statistical modelling approach. Sahana *et al* (2018) use the so-called Least Absolute Shrinkage and Selection Operator to select the most significant grids of predictors [14]. Stolbova *et al* (2016) propose a method to select key regions for prediction that is based on complex networks [15]. While it is difficult to systematically compare the prediction skills of previous works due to differences in the definition of onset dates, prediction years, as well as lead times, the most successful statistical models provide prediction results with a root mean square error (RMSE) of about 4 days. For example, Terzi *et al* (2019) predicted the

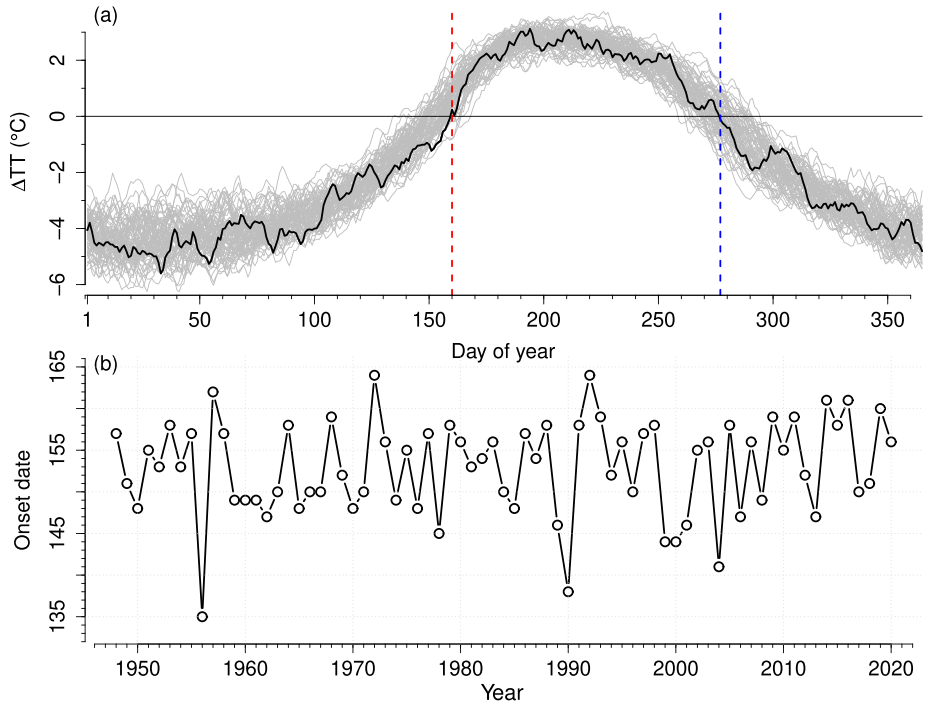

Figure 1. (a) The meridional gradient of the tropospheric temperature ΔTT as a function of the day of the year: 2019 (Black) and other years (grey) from the time span from 1948 to 2020. The red and blue dashed lines show the onset and withdrawal dates of 2019, respectively. (b) The onset dates of the ISM based on the ΔTT index over 1948–2020. The *y*-axis is the onset date counted from January 1st of each year. The mean onset date is June 2nd (153rd day), with a standard deviation of 5.8 days.

monsoon onset over Kerala (MOK) over 2009–2018 with a RMSE of 3.7 days, with an average lead time of 52 days, by using monitored Beryllium-7 time series [12]. Numerical weather models have been shown to have high skill for predicting the ISM onset dates [4, 14, 16, 17]. The UK Met Office Global Seasonal Forecasting system, GloSea5-GC2, initialized on April 25th (i.e. with an average lead time of around 38 days) provides an ensemble of predictions with mean correlation of 0.8 for onset dates defined with the tropospheric temperature gradient ΔTT [17]. The ECMWF Seasonal Forecasting System 4, initialized with ocean, atmosphere and snow conditions on April 1st, provides an ensemble prediction with mean correlation of 0.77 for ΔTT [18], at a lead time with respect to the mean onset date (May 29th in their study) of 58 days.

In both statistical and numerical modelling approaches, typical lead times are within the extended to subseasonal range of 10–60 days. Seasonal predictions with a lead time longer than two months are rare. Exceptionally, Pradhan *et al* (2017) predicted the onset dates based on ΔTT index over 1982–2008 by using global seasonal forecasting models (CFSv2-T126 and CFSv2-T382), three months in advance [4]. While the authors succeed in predicting 70% of early onsets and 60% of delayed onsets, the RMSE of their prediction is larger than 6.5 days and the correlation is limited to 0.49 [4]. It is therefore of great interest to investigate the seasonal predictability of the ISM onset date further.

Neural networks have recently been shown to have at least comparable forecast skill compared to numerical, process-based models for short-term (up to three days) to medium-range (up to two weeks) weather forecast [19–21]. A deep learning architecture has recently been shown to outperform numerical forecast models for annual-scale prediction of the El Niño Southern Oscillation (ENSO) [22]. Moreover, a deep learning approach has been demonstrated to accurately reconstruct spatial fields of surface temperatures [23]. In this paper, we use an echo state network (ESN), which is a sparsely and randomly connected recurrent neural network with input and output layers [24, 25] (see figure S2 for a schematic) for seasonal prediction of the ISM onset. The ESN transforms a set of input signals, $\mathbf{u}(t)$, into high-dimensional temporal patterns of the nodes $\mathbf{r}(t)$ and exploits those patterns to predict a target signal $s(t)$. Only the weights of output layer are optimized to approximate the target signal, for which a regularized linear regression method is employed. This type of machine learning is referred to as reservoir computing [24–28] and has the advantage of comparably simple and very efficient training. Reservoir computers such as ESNs have been shown to be useful in a wide range of applications, including prediction [29–31] and partial inferences [32–34] of time series from chaotic and complex systems, speech recognition [35, 36], or control of robotic systems [28, 37]. In the field of climate science, Huang *et al* recently showed that the tropical average surface

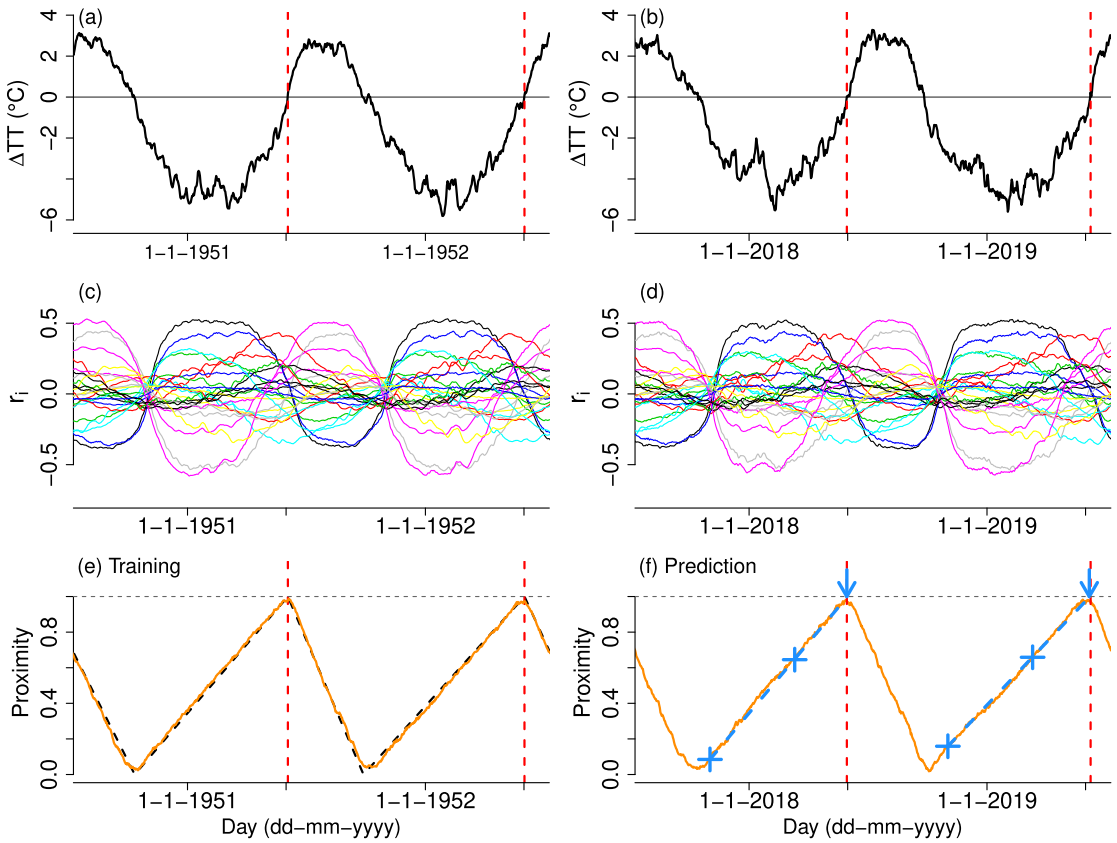

Figure 2. Prediction scheme of the onset dates of the Indian Summer Monsoon with echo state networks (ESNs). (a), (b) The daily series of the meridional gradient of the tropospheric temperature ΔTT for 1951 (a year during the training period) and 2018 (a year during the test period). The red dashed lines show the onset dates, objectively defined as the first days at which ΔTT crosses zero in each year. (c), (d) The dynamics of ESN nodes $r_i(t)$ (only $1 \leqslant i \leqslant 25$ are shown for illustration) corresponding to the two time intervals shown in (a) and (b), respectively. (e) The proximity function defined from the ΔTT index (black dashed). The proximity function estimated with the ESN (orange solid) for the training period. (f) The proximity function estimated with the ESN (orange solid) for the test period. Cross marks show the reference time points t_1 taken as 60 days before January 1st of a prediction year (November 2nd), and t_2, taken as the 70th day of the prediction year (March 10th in non-leap years). The dashed blue lines show the linear extrapolations using the reference points t_1 and t_2. The arrows indicate the predicted onset dates, where the extrapolation reaches +1.

air temperature can be reconstructed well from the average Northern Hemisphere surface air temperature over 1981–2018 by using an ESN; for this task, the ESN outperforms a traditional backpropagation-based neural network or the long short-term memory neural network [34].

We propose a prediction method of the ISM onset date based on ESNs, which we outline as follows (see also figure S2). We first introduce a function of time that takes 1 at the ISM onset dates and 0 at the withdrawal dates, and is linear between these points. A value of the function indicates the temporal proximity to an upcoming onset (withdrawal) when the function is increasing (decreasing). We call this function the *proximity function*, and use it as the teacher signal for the supervised learning of ESNs. Over a training period, an ensemble of ESNs are trained to approximate the proximity function on the basis of two input signals, the ΔTT and a sinusoidal seasonal forcing. The trained ESNs are then used to predict the proximity function over the prediction period. If the proximity function is accurately estimated up to a date t_2 before the actual ISM onset, the onset date t_{on} can be predicted in advance based on the linear trend of the estimated proximity function between a reference time t_1 and the later t_2 (that is, by linear extrapolation, see figure 2(f)). Finally, an ensemble prediction of onset dates is performed.

The paper is structured as follows: in section 2 we describe the employed data and explain the developed methodology. The results are shown in section 3 and discussed in section 4, which ends with a brief summary of the main findings. Some further details are provided in supplementary information (available online at stacks.iop.org/ERL/16/074024/mmedia).

2. Data and method

2.1. Data

There are various definitions of ISM onset and withdrawal dates, based on different physical variables (e.g. precipitation, wind speed, etc) and spatial scales (from local to continental) [3, 4, 13, 38–43]. Since 2006 the Indian Meteorological Department

(IMD) has reported the date of the MOK, objectively defined with rainfall over 14 stations in Kerala, wind fields, and outgoing long wave radiation [13]. There are multiple other onset indices defined based on large-scale circulation patterns or thermodynamic conditions. The most widely-used indices are the ΔTT-based index [2, 3], the hydrological onset and withdrawal index [39], and the onset circulation index [41]. These large-scale indices, which result from averaging over extensive spatial regions, are less susceptible to 'false' or 'bogus' onsets due to synoptic disturbances or intraseasonal oscillations. In this work we focus on the ΔTT index, which is the meridional gradient of tropospheric temperature over the southern Euracian region [3, 44]. Specifically, ΔTT is calculated as the difference between the average tropospheric temperatures over 600–200 hPa in a northern box (30°E–110°E; 5°N–35°N) and a southern box (30°E–110°E; 15°S–5°N) [4]; see figure S1. Tropospheric temperature fields are obtained from the NCEP/NCAR daily reanalysis fields (1948–2020) [45] as in previous works [2–4]. While the NCEP/NCAR reanalysis data is not of the highest resolution, it has been shown to be more robust against bogus onsets than ERA-40 [2] and, crucially, it enables us to carry out a straightforward comparison to the recent work by Pradhan *et al* (2017) [4]. The change of sign of ΔTT indicates the shift of the deep tropospheric heat source that drives the ISM circulation, form south to north and from north to south, respectively [2, 3]. Thus, an onset (withdrawal) date of the ISM is defined as the day at which ΔTT changes its sign from negative to positive (positive to negative) as shown in figure 1(a). The ΔTT-based onset dates are well correlated with the objectively-defined dates of MOK [13] over 1971–2020 ($r = 0.57$, figure S9). The average over 1971–2020 is June 2nd (∼153rd day of year) for both definitions. The standard deviation (s.d.) of the ΔTT-based onset dates is 5.8 days, which is slightly smaller than that of MOK, for which the s.d. is 6.6 days over 1971–2020.

2.2. Echo state networks (ESN)

We outline our specifications of ESNs [24, 25] essentially following Lu *et al* (2017) [32]. The ESN consists of N randomly connected nodes, and has one input layer and one output layer (see schematic in figure S2). Node i has a time-dependent state $r_i(t)$, and the whole state of the ESN is specified by the vector $\mathbf{r}(t) = (r_1(t), r_2(t), .., r_N(t))^T \in \mathbf{R}^N$. At every time step, the ESN receives a set of M inputs $\mathbf{u}(t) = (u_1(t), u_2(t), .., u_M(t))^T \in \mathbf{R}^M$, and the state evolves according to:

$$\mathbf{r}(t+1) = \tanh(A\mathbf{r}(t) + W_{\text{in}}\mathbf{u}(t) + \xi 1),$$

where $A \in \mathbf{R}^{N\times N}$ is the weighted adjacency matrix of the network, $W_{\text{in}} \in \mathbf{R}^{N\times M}$ is the input weight matrix, 1 is a vector of ones, and ξ is a bias parameter. The adjacency matrix A is built from a sparse random Erdös-Rényi matrix in which the fraction of nonzero matrix elements is D/N, so that the average degree of a node is D. The values of non-zero elements are drawn independently from the uniform distribution over $[-1, 1]$. We then uniformly rescale all the elements of A (i.e. multiply A by a positive scalar) such that the largest value of the magnitudes of its eigenvalues equals a predefined scalar ρ, which we refer to as the spectral radius of A. For the input layer, the ith of the M input signals is connected to all nodes with connection weights in the ith column of W_{in}. The elements of W_{in} are randomly chosen from the uniform distribution over $[-\sigma, \sigma]$ for another scalar parameter σ. The bias parameter ξ may help to overcome undesired consequences of the symmetry of the tanh function with respect to 0 and is commonly used.

In this study we consider ESNs with a single, scalar output:

$$\hat{s}(t) = W_{\text{out}}\mathbf{r}(t) + c,$$

where $W_{\text{out}} \in \mathbf{R}^{1\times N}$ is the output weight matrix and c is a constant. For a teacher signal $\{s(t)|t = 1, ..., T\}$, the optimal output weights W^*_{out} and constant c^* are given so that a quadratic form:

$$\left\{\sum_{t=1}^{T} |W_{\text{out}}\mathbf{r}(t) + c - s(t)|^2\right\} + \beta[\text{Tr}(W_{\text{out}}W^T_{\text{out}})],$$

is minimized. The second term $\beta[\text{Tr}(W_{\text{out}}W^T_{\text{out}})]$ is a regularization term to avoid over-fitting, and β is typically chosen to be small positive number called the ridge regression parameter. This is also sometimes referred to as Tikhonov regularization. The solution of this minimization problem is known to be:

$$\begin{aligned} W^*_{\text{out}} &= \delta\mathbf{S}\delta\mathbf{R}^T(\delta\mathbf{R}\delta\mathbf{R}^T + \beta\mathbf{I})^{-1}, \\ c^* &= -[W^*_{\text{out}}\bar{\mathbf{r}} - \bar{s}], \end{aligned}$$

where $\bar{\mathbf{r}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{r}(t)$, $\bar{s} = \frac{1}{T}\sum_{t=1}^{T}s(t)$, $\mathbf{I}$ is the identity matrix, and $\delta\mathbf{R}$ (respectively $\delta\mathbf{S}$) is the matrix whose kth column is $\mathbf{r}(t) - \bar{\mathbf{r}}$ (respectively $s(t) - \bar{s}$) [32].

For best performance, ESNs should have the so-called *echo-state property* [25] (or in other words *consistency* [31]), which essentially states that the effect of a previous state $\mathbf{r}(t)$ and a previous input $\mathbf{u}(t)$ vanishes gradually in the future state $\mathbf{r}(t+k)$ as time passes (i.e. $k \to \infty$). This property is achieved by suitably setting the spectral radius of the matrix A. For most practical purposes, the echo state property is assured if the spectral radius ρ of the network matrix A is less than unity [25]. On the other hand, larger ρ also has the effect of driving signals $\mathbf{r}(t)$ into more nonlinear regions of tanh units. As a rule of thumb, ρ should be close to 1 for tasks that require long memory and accordingly smaller for tasks where a too long memory might in fact be harmful. Thus in

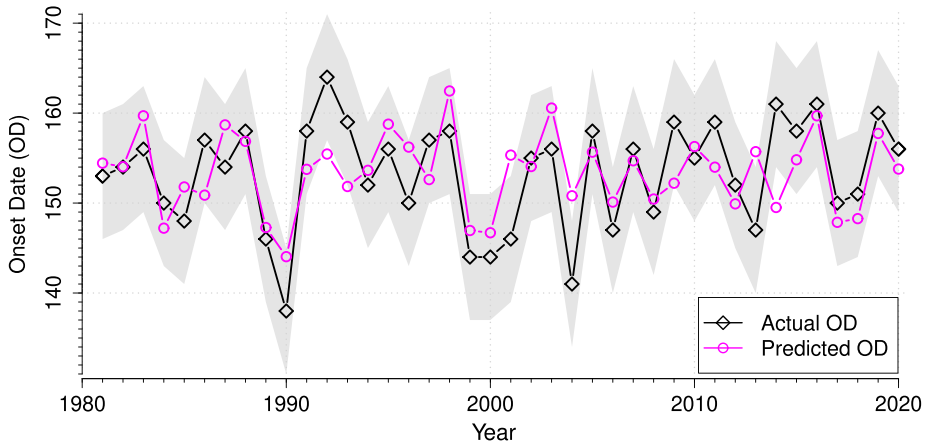

Figure 3. Prediction results of the onset dates of the Indian Summer Monsoon over 1981–2020. The actual onset date is calculated based on the sign of the meridional gradient of tropospheric temperature ΔTT (see section 2). The predicted onset date is the ensemble mean of predictions of the top 100 ESNs (see text). The shade shows the range of ±7 days with respect to the actual onset date from ΔTT data. Statistical summaries are given in table S1.

this work, we choose $\rho = 1$ and the other parameters as $N = 100$, $D = 5$, $\sigma = 0.04$, $\xi = 0.01$ and $\beta = 10^{-10}$. The results are robust against small changes in these parameters.

3. Results

While other definitions exist, we focus on the objective definition of the ISM onset and withdrawal dates via the meridional tropospheric temperature gradient ΔTT [3, 4, 44] for the time interval 1948–2020 (figure 1). The daily series of ΔTT is calculated from NCEP/NCAR reanalysis temperature [45] (see section 2). As mentioned above, Pradhan *et al* (2017) predicted the ΔTT-based onset dates for the period 1982–2008 by using global seasonal forecasting models three months in advance [4]. We predict the onset dates over the 40 years from 1981 to 2020 and compare the skill over 1982–2008 with that of Pradhan *et al* (2017).

The onset dates over 1981–2020 are predicted by training ESNs using the onset and withdrawal dates observed over 1948–1980. We can *a priori* define the proximity function between the onset date in 1948 and the withdrawal date in 1980. The proximity function is shown by the dashed black line in figure 2(e). We generate 10^5 ESNs with different, randomly chosen weights and connections for the purpose of an ensemble forecast (see section 2). A node of each ESN receives two input signals: one is $\Delta\mathrm{TT}(t)$ and the other is a seasonal forcing $\cos(2\pi(t - 152.25)/365.25)$, where t is a day counted from January 1st 1948, and 152.25 is the average day of the year corresponding to June 1st (taking into account leap years). These input signals are transformed into temporal patterns of ESN nodes (figures 2(c) and (d)). The output layer of each ESN is optimized to fit the proximity function over a training period between the onset date in 1950 and the onset date in 1980 (the years 1948–1949 are excluded from the training period because the ESN states can be affected by the initial conditions). The estimated proximity function is shown for 1951 in figure 2(e) in solid orange. In the following prediction period after the onset date in 1980, the proximity function is estimated with the trained ESNs. An onset date is then predicted by linear extrapolation using the values of the estimated proximity function on two reference dates t_1 and t_2, as shown in figure 2(f). Unless otherwise mentioned, t_1 is the date 60 days before January 1st of a prediction year (i.e. November 2nd) and t_2 is the 70th day of the prediction year (March 10th in non-leap years). The lead time for the mean onset date is therefore $83 (= 153 - t_2)$ days. We conduct an ensemble forecast since the predicted onset dates depend on the random realizations of ESNs, and in order to provide forecast uncertainties. The prediction of the onset dates by each ESN is statistically more accurate if the ESN has a lower RMSE between the estimated and actual proximity function over the training period (figure S3). Therefore, we select the top 100 ESNs, i.e. the 100 ESN realizations with the lowest RMSEs on the training period, and then take the ensemble mean of the predicted onset dates of these 100 ESNs as the final prediction.

Figure 3 compares the predicted onset dates with the actual ones over the time span 1981–2020, where the RMSE is 4.9 days and the correlation between the predicted and actual yearly onset date time series is 0.58 (table S1). The normalized RMSE (nRMSE), namely the RMSE divided by the standard deviation of the observed onset dates, is 0.82. For the shorter time span 1982–2008, the skill scores are 4.8 days (RMSE), 0.76 (nRMSE) and 0.67 (correlation) (table S1). It should be noted that the standard deviation of the predicted onset dates is 30% smaller than

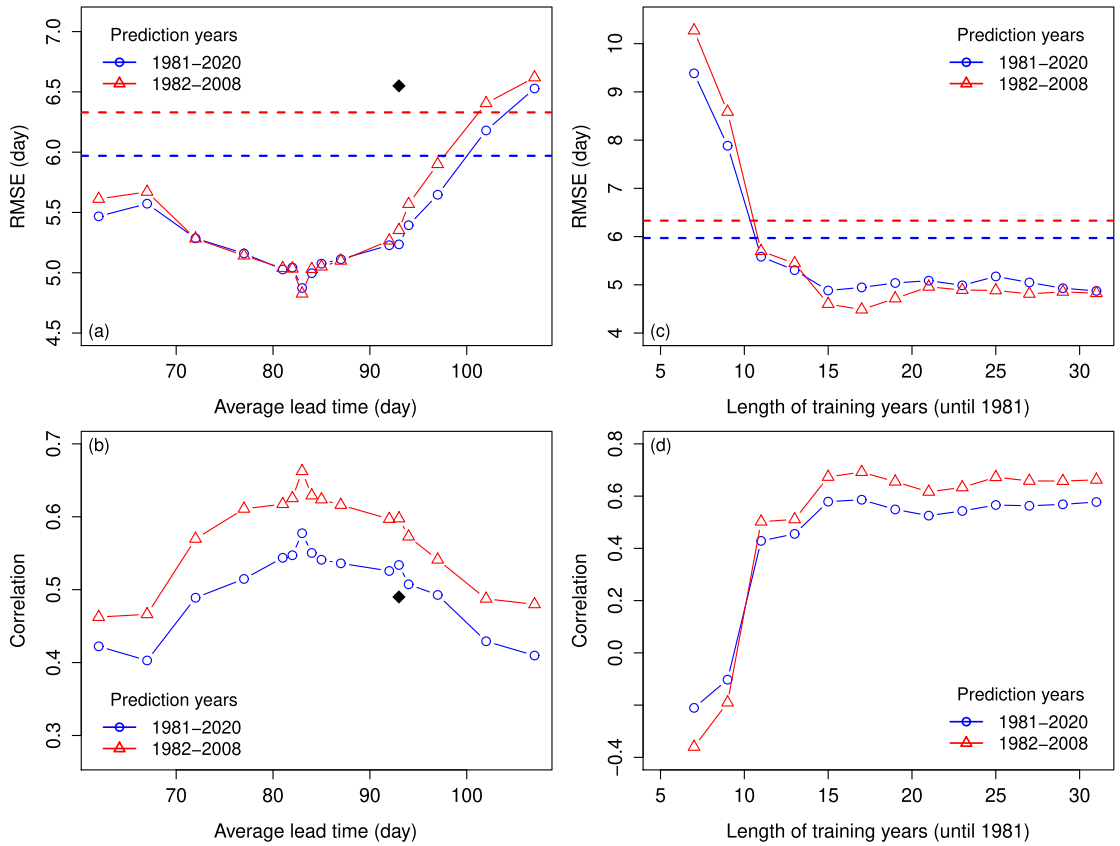

Figure 4. Dependence of prediction skills on the average lead time (left) and the length of training period (right). The average lead time is the number of days between the issuing of the forecast at t_2 and June 2nd (the average onset date of data). The dashed horizontal lines show the standard deviations (s.d.) of the onset dates of data in each prediction period (table S1). The black diamonds show the results reported by Pradhan *et al* [4]. In the experiments in the right panels, the training interval is set to end in 1980. For the left panels, the training length is 31 years from 1950 to 1980. For the right panels, the average lead time is 83 days.

that of the actual onset dates. The distributions of the ensemble predictions of the 100 ESNs are shown in figure S4.

The above results are robust against moderate changes of lead time by ±10 days (figures 4(a) and (b)). For the case with the average lead time of three months (93 days) used by Pradhan *et al* (2017), we obtain the skill 5.35 days (RMSE), 0.84 (nRMSE) and 0.60 (correlation) over the period 1982–2008, compared to 6.55 days (RMSE), 0.88 (nRMSE) and 0.49 (correlation) obtained by the seasonal forecasting model by Pradhan *et al* (2017) [4] (figures 4(a) and (b) and table S2). Maybe surprisingly, the skill moderately decreases as the lead time gets shorter than 83 days. The reason is that it is more difficult for ESNs to approximate the proximity function near onset dates, where the function bends sharply. We also compare our predicted onset dates with the dates of the MOK reported by the IMD [13] (see section 2 and figure S6). As shown in table S3, the skill scores with respect to the MOK are slightly worse than those with respect to the large-scale ΔTT-based onset dates. This is expected because our ESNs are optimized on the ΔTT-based onset dates rather on than the dates of MOK, and can hence by construction not capture differences between the large-scale onset dates and the local MOK dates. Nevertheless, our predictions are significantly correlated with MOK, and the RMSE is below the standard deviation of the dates of MOK (table S3).

We have performed sensitivity experiments on the choice of the training interval and found that the skill scores remain similar over 1981–2020 as well as 1982–2008 (figure 4). The result shows that a training period of length 15 years is sufficient to obtain skilful predictions. We have also examined a different way of setting training and prediction years, where the onset date of each year is predicted by using the onset and withdrawal dates of previous 31 years (for example, the onset date in 2020 is predicted with the onset and withdrawal dates over 1988–2019). This scheme provides the prediction skill of 5.19 days (RMSE) and 0.53 (correlation) over 1981–2020, which are only slightly worse than those for the setup described above (table S1). We conclude that our prediction scheme is robust against the choice of the training interval. We note that the sinusoidal input is necessary to obtain a skilful forecast, likely because it reduces estimation errors of the proximity function by preventing ESNs from overweighting local fluctuations in the ΔTT signal (compare figure S7 with figure 2).

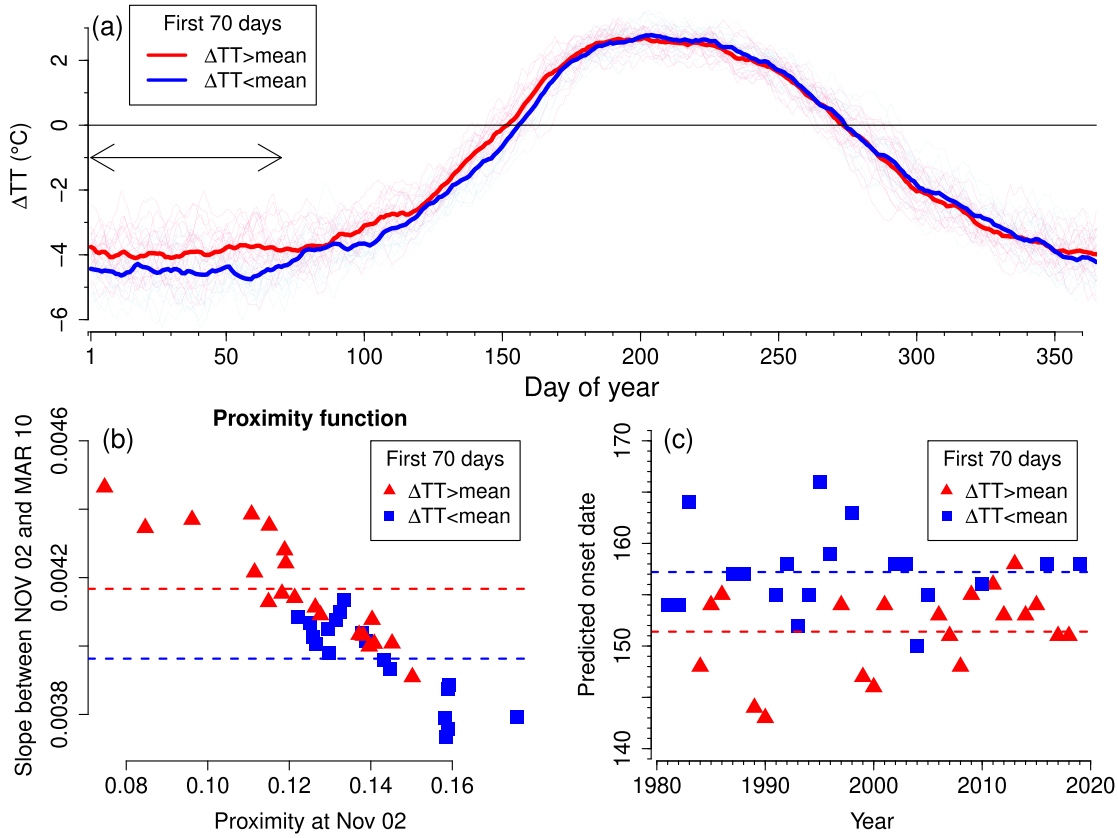

Figure 5. (a) The meridioal tropospheric temperature gradient ΔTT as functions of the day of the year. The ΔTT signals with above-average winter ΔTT are shown in pink and those with below-average winter ΔTT are shown in skyblue, respectively, where the winter ΔTT refers to the mean value of ΔTT over the first 70 days of the year. The thick red (blue) lines show the mean of the ΔTTs with higher (lower) winter ΔTT. When mean winter ΔTTs are high, onset dates are earlier on average, and vice versa. (b) The scatter plot of the slope of the proximity function and its value at November 2nd for a typical ESN. When the ΔTT is higher in winter, the slope of the proximity function between two reference dates (November 2nd and March 10th) tends to be steeper. The red and blue dashed lines show respective means. (c) A steeper slope of the proximity function leads to a prediction an earlier onset on average.

By scrutinizing variations of ΔTT and ISM onset dates (figure 5(a)), we find that a higher (lower) boreal winter ΔTT typically results in an earlier (later) onset date (the ΔTT averaged over the first 70 days of the year is correlated with the ISM onset date with a correlation of −0.52 over 1981–2020). Hence the boundary conditions determining the ISM onset in terms of the large-scale tropospheric temperature gradient are already developed during the preceding winter season. In our approach, a higher winter ΔTT is translated into a proximity function with a higher slope, leading to the prediction of an earlier onset, and vice versa (figures 5(b) and (c)). We note here that our nonlinear ESN approach provides a more accurate prediction of the ISM onset dates than a linear regression using the winter ΔTT as the explanatory variable (figure S8).

We finally discuss a possible factor which affects the seasonal predictability of the monsoon onset. Figure S5 shows the time evolution of the ΔTT index for nine years for which our method results in comparably weak predictions. For five of these nine years, the evolution of the ΔTT stagnates or fluctuates in the vicinity of the threshold ΔTT = 0 that defines the ISM onset (specifically years 1986, 1993, 1996, 2004, and 2009 in figure S5). In these cases, small fluctuations in the ΔTT near the threshold could yield a large difference in the onset date. An accurate prediction is difficult in such cases, where the definition of the target is already uncertain due to the noise in the ground truth data. We have not found other clear factors lowering the predictability (see caption of figure S5 for more discussions, including effects of ENSO).

4. Summary and discussion

It is known that the ISM onset dates depend on multiple interannual climate phenomena such as ENSO [2, 3], the Indian Ocean Dipole [5, 46], the North Atlantic Oscillation [47], as well as the Himalayan–Tibetan Plateau snowpack [18, 48]. These factors aggregately affect the ΔTT anomaly [2, 3, 5, 18, 44] and hence the ISM onset dates at seasonal time scales. Our method thus indirectly exploits the large-scale teleconnections induced by these phenomena via the ΔTT evolution to achieve a skilful prediction of the ISM onset at comparably long forecast horizons.

We introduced a neural-network-based method for predicting real-world nonlinear phenomena such as abrupt state transitions in the climate system, for which only short and noisy observational training data are available. We applied our method to predict the ISM onset dates, which were objectively defined by the meridional gradient of the tropospheric temperature ΔTT index. In our proximity inference method, ESNs are used to infer the temporal proximity to the ISM onset from the time series of the ΔTT index. Our approach enables us to predict the ISM onset dates with seasonal-scale lead times, with an RMSE ranging from 4.5 to 5.4 days and a correlation coefficient ranging from 0.5 to 0.7, even when only 15 years of training data are used (figures 4(c) and (d)). Our method hence extends the forecast horizon of previous studies from subseasonal to seasonal time scales at comparable skill, and outperforms a state-of-the-art numerical model prediction [4] at comparable lead times. To our knowledge, there is no statistical forecasting approach at comparable, seasonal scales. Our method also outperforms the statistical forecast introduced by Stolbova *et al* [15] in terms of their accuracy test, which considers a forecast to be valid if the predicted onset date is within ±7 days of the actual onset date: for the common time span 1981–2015, we obtain an accuracy of 80% compared to 71%, despite the 1.5 months longer lead time. However, it should be noted that the definition of onsets are different in these two studies, as Stolbova *et al* focus on the onsets in the Eastern Ghats [15]. In comparison to the ensemble forecast that we focused on here, a single ESN can lead to forecasts with lower RMSEs. For example, for the test period over 1981–2020, we obtain an RMSE of about 4.4 days with the best-performing ESN, as shown in figure S3.

Due to the stable performance for different choices of training and test periods within the time span for which data is available, we expect our approach to exhibit similar skill also for future predictions. In this work we focused on the prediction of the onset dates of the large-scale ISM based on the ΔTT index. Nevertheless, our model trained on the ΔTT-based onset dates has a certain skill for predicting the dates of the MOK as well: the correlation is 0.47 and the RMSE is 5.6 days, which is lower than the standard deviation of MOK (6.2 days, see table S3 and figure S6). This suggests a potential of extending the present method to the prediction of locally defined onset dates as well. It is straightforward to apply our method for other similar threshold-based indices such as the Onset Circulation Index [41] or the Hydrologic Onset and Withdrawal Index [39]. The predictability for these indies will be assessed in future work. We will also address large-scale impacts on the ISM onset and overall rainfall sums, such as connections to ENSO [2, 3, 5, 6], the Pacific Decadal Oscillation [49], the Indian Ocean Dipole [5, 46], and the North Atlantic Oscillation [47], as well as extensions of our approach to predict break phases [50] of the ISM, in future research.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://github.com/takahito321/Indian_monsoon/tree/master.

The NCEP-NCAR reanalysis daily temperature data used in this study can be obtained from https://psl.noaa.gov/data/gridded/data.ncep.reanalysis.html. We have used a free software environment R for statistical computing and graphics. Our codes are available at GitHub repository: https://github.com/takahito321/Indian_monsoon/tree/master.

Acknowledgments

T M and N B acknowledge funding by the Volkswagen Foundation. This project is TiPES contribution #120: The TiPES ('Tipping Points in the Earth System') project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 820970.

Author contributions

T M and N B conceived the study. T M conducted the analysis. Both authors interpreted and discussed the results and wrote the manuscript.

Competing interests

The authors declare no competing interests.

ORCID iDs

Takahito Mitsui https://orcid.org/0000-0002-2825-3996
Niklas Boers https://orcid.org/0000-0002-1239-9034

References

[1] Kumar K N, Rajeevan M, Pai D S, Srivastava A K and Preethi B 2013 On the observed variability of monsoon droughts over India. *Weather. Clim. Extremes* **1** 42–50
[2] Goswami B and Xavier P K 2005 *Geophys. Res. Lett.* **32** 1–4
[3] Xavier P K, Marzin C and Goswami B 2007 *Q. J. R. Meteorolog. Soc.* A **133** 749–64
[4] Pradhan M, Rao A S, Srivastava A, Dakate A, Salunke K and Shameera K 2017 *Sci. Rep.* **7** 1–14
[5] Sabeerali C, Rao S A, Ajayamohan R and Murtugudde R 2012 *Clim. Dyn.* **39** 841–59
[6] Webster P J, Magana V O, Palmer T, Shukla J, Tomas R, Yanai M and Yasunari T 1998 *J. Geophys. Res.: Oceans* **103** 14451–510
[7] Jain M, Naeem S, Orlove B, Modi V and DeFries R S 2015 *Glob. Environ. Change* **31** 98–109
[8] Kung E C and Sharif T A 1982 *J. Meteorolog. Soc. Japan. Ser. II* **60** 672–81

[9] Kumar M R 2004 *IEEE Geosci. Remote Sens. Lett.* **1** 265–7
[10] Ancy S, Kumar R, Asokan R and Subhashini R 2014 Prediction of onset of south west monsoon using multiple regression *Proc. IEEE Int. Conf. Computer Communication and Systems ICCCS14* (IEEE) pp 170–5
[11] Ajitha T and Kumar P A 2017 *J. Trop. Agric.* **55** 31–9
[12] Terzi L, Kalinowski M, Schoeppner M and Wotawa G 2019 *Sci. Rep.* **9** 1–6
[13] Pai D and Nair R M 2009 *J. Earth Syst. Sci.* **118** 123–35
[14] Sahana A and Ghosh S 2018 *Geophys. Res. Lett.* **45** 8510–18
[15] Stolbova V, Surovyatkina E, Bookhagen B and Kurths J 2016 *Geophys. Res. Lett.* **43** 3982–90
[16] Alessandri A, Borrelli A, Cherchi A, Materia S, Navarra A, Lee J Y and Wang B 2015 *Mon. Weather Rev.* **143** 778–93
[17] Chevuturi A, Turner A G, Woolnough S J, Martin G M and MacLachlan C 2019 *Clim. Dyn.* **52** 6599–617
[18] Senan R, Orsolini Y J, Weisheimer A, Vitart F, Balsamo G, Stockdale T N, Dutra E, Doblas-Reyes F J and Basang D 2016 *Clim. Dyn.* **47** 2709–25
[19] Rasp S, Dueben P D, Scher S, Weyn J A, Mouatadid S and Thuerey N 2020 *J. Adv. Model. Earth Syst.* **n/a** e2020MS002203
[20] Rasp S and Thuerey N 2020 (arXiv:2008.08626)
[21] Sønderby C K, Espeholt L, Heek J, Dehghani M, Oliver A, Salimans T, Agrawal S, Hickey J and Kalchbrenner N 2020 arXiv:2003.12140
[22] Ham Y G, Kim J H and Luo J J 2019 *Nature* **573** 568–72
[23] Kadow C, Hall D M and Ulbrich U 2020 *Nat. Geosci.* **13** 408–13
[24] Jaeger H 2001 *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report* 148 13
[25] Lukoševičius M and Jaeger H 2009 *Comput. Sci. Rev.* **3** 127–49
[26] Maass W, Natschläger T and Markram H 2002 *Neural Comput.* **14** 2531–60
[27] Tanaka G, Yamane T, Héroux J B, Nakane R, Kanazawa N, Takeda S, Numata H, Nakano D and Hirose A 2019 *Neural Netw.* **115** 100–23
[28] Nakajima K 2020 *Japan. J. Appl. Phys.* **59** 060501
[29] Pathak J, Hunt B, Girvan M, Lu Z and Ott E 2018 *Phys. Rev. Lett.* **120** 024102
[30] Jaeger H and Haas H 2004 *Science* **304** 78–80
[31] Inubushi M and Yoshimura K 2017 *Sci. Rep.* **7** 1–10
[32] Lu Z, Pathak J, Hunt B, Girvan M, Brockett R and Ott E 2017 *Chaos* **27** 041102
[33] Nakai K and Saiki Y 2018 *Phys. Rev.* E **98** 023111
[34] Huang Y, Yang L and Fu Z 2020 *Earth Syst. Dyn.* **11** 835–53
[35] Verstraeten D, Schrauwen B and Stroobandt D 2005 Isolated word recognition using a liquid state machine *ESANN* vol 5 (Citeseer) pp 435–40
[36] Paquot Y, Duport F, Smerieri A, Dambre J, Schrauwen B, Haelterman M and Massar S 2012 *Sci. Rep.* **2** 287
[37] Salmen M and Ploger P G 2005 Echo state networks used for motor control *Proc. 2005 IEEE Int. Conf. Robotics and Automation* (IEEE) pp 1953–8
[38] Webster P J and Yang S 1992 *Q. J. R. Meteorol. Soc.* **118** 877–926
[39] Fasullo J and Webster P 2003 *J. Clim.* **16** 3200–11
[40] Taniguchi K and Koike T 2006 Comparison of definitions of Indian summer monsoon onset: Better representation of rapid transitions of atmospheric conditions *Geophys. Res. Lett.* **33**
[41] Wang B, Ding Q and Joseph P 2009 *J. Clim.* **22** 3303–16
[42] Noska R and Misra V 2016 *Geophys. Res. Lett.* **43** 4547–54
[43] Walker J M and Bordoni S 2016 *Geophys. Res. Lett.* **43** 11–815
[44] Goswami B, Wu G and Yasunari T 2006 *J. Clim.* **19** 5078–99
[45] Kalnay E *et al* 1996 *Bull. Am. Meteorol. Soc.* **77** 437–72
[46] Ashok K, Guan Z and Yamagata T 2001 *Geophys. Res. Lett.* **28** 4499–502
[47] Goswami B N, Madhusoodanan M, Neema C and Sengupta D 2006 *Geophys. Res. Lett.* **33**
[48] Blanford H F 1884 *Proc. R. Soc.* A **37** 3–22
[49] Krishnan R and Sugi M 2003 *Clim. Dyn.* **21** 233–42
[50] Krishnan R, Zhang C and Sugi M 2000 *J. Atmos. Sci.* **57** 1354–72