COVID-19 Time Series Forecast Using Transmission Rate and Meteorological Parameters as Features

Mohsen Mousavi
University of Tasmania, AUSTRALIA and University of Technology Sydney, AUSTRALIA

Rohit Salgotra
Thapar Institute of Engineering & Technology, INDIA

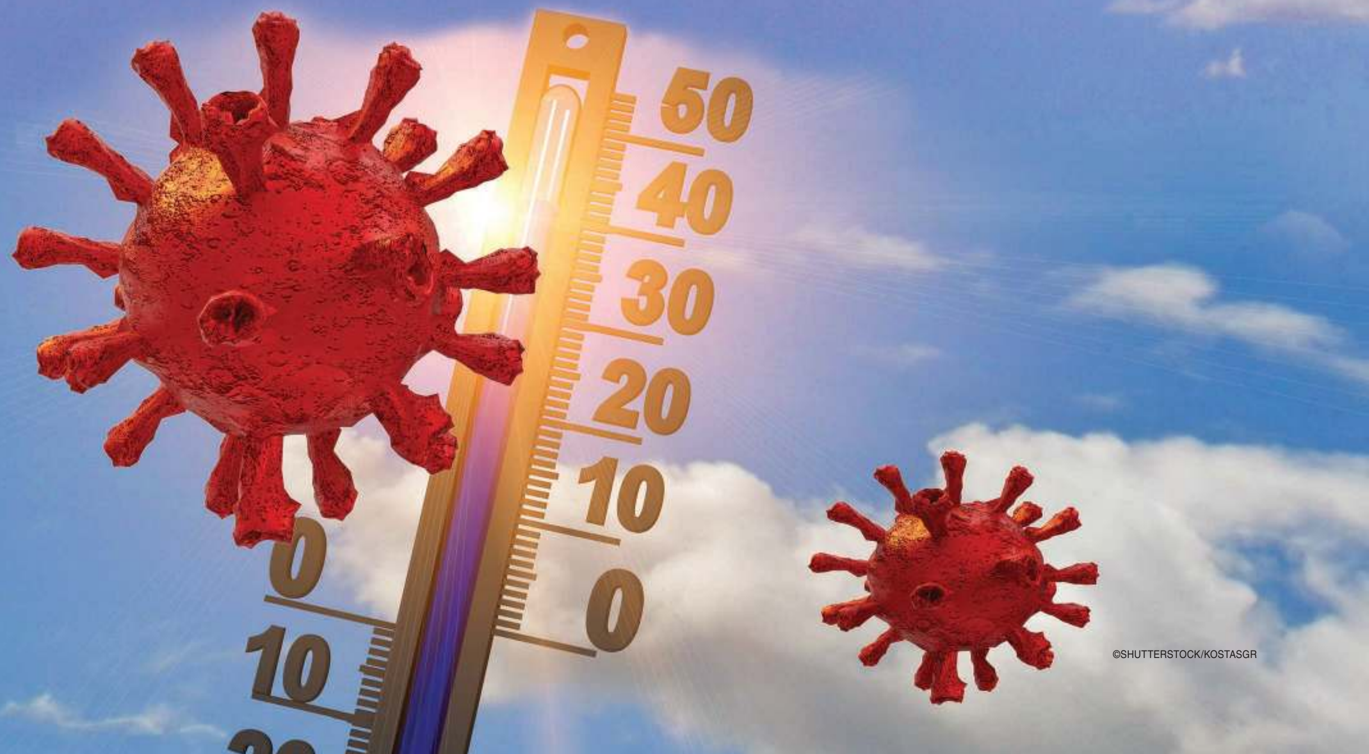
Damien Holloway
University of Tasmania, AUSTRALIA

Amir H. Gandomi
University of Technology Sydney, AUSTRALIA

Digital Object Identifier 10.1109/MCI.2020.3019895
Date of current version: 14 October 2020

Abstract—The number of confirmed cases of COVID-19 has been ever increasing worldwide since its outbreak in Wuhan, China. As such, many researchers have sought to predict the dynamics of the virus spread in different parts of the globe. In this paper, a novel systematic platform for prediction of the future number of confirmed cases of COVID-19 is proposed, based on several factors such as transmission rate, temperature, and humidity. The proposed strategy derives systematically a set of appropriate features for training Recurrent Neural Networks (RNN). To that end, the number of confirmed cases (CC) of COVID-19 in three states of India (Maharashtra, Tamil Nadu and Gujarat) is taken as a case study. It has been noted that stationary and non-stationary parts of the features improved the prediction of the stationary and non-stationary trends of the number of confirmed cases, respectively. The new platform has general application and can be used for pandemic time series forecasting.

Corresponding Author: Amir H. Gandomi (gandomi@uts.edu.au).



©SHUTTERSTOCK/KOSTASGR

I. Introduction

The novel coronavirus (SARS-CoV-2) has plunged the world into severe disaster recently. The virus made its way to many countries around the globe soon after the first case was reported in Wuhan, Hubei Province, People's Republic of China (PRC) in late December [1]. As such, the World Health Organization (WHO) declared the situation as a public health emergency of international concern on 30 January, 2020 [2]. WHO officially named the disease COVID-19 when PRC Center for Disease Control and Prevention (CDC) recognized the virus as a new type of coronavirus. Ever since, many countries have experienced disasters due to the widespreading infectious virus. This has put a huge burden on medical centers in different countries and many different measures have been put in place by jurisdictions to control the spread of the virus in different countries. These measures are mainly in the form of lockdowns enforced in several stages, where people are banned from congregating *en messe*. Physical distancing measures can have a huge impact on the virus transmission rate [3], and in one such study the time taken for the daily number of new cases to double was reported to increase from 2 to 4 days [4]. The optimal lockdown policy depends on the fraction of infected and susceptible in the population. As a result, a severe lockdown beginning two weeks after the outbreak was prescribed where it can be gradually relaxed to cover 60% of the population after a month, and 20% of the population after three months [5]. It was also recommended that the intensity of the lockdown should depend on the gradient of the fatality rate as a proportion of the infected, and on the assumed value of a statistical life [5].

... in this section we propose a novel framework for optimum training of an RNN using time series analysis of the signals used in training.

The effect of the Meteorological parameters on the spread of the COVID-19 disease has also been investigated [6], [7]. It has been reported that the virus favors low temperature and low humidity [8]–[11]. Mortality is also shown to be affected by temperature and humidity variation [12]: one unit increase of temperature and absolute humidity was associated with a decreased COVID-19 death rate. Accordingly, temperature and humidity are suggested as important factors to be considered in modelling of rates.

Some studies have focused on prediction of the number of future cases with different lockdown policies in different countries [13]–[15]. This will facilitate the investigation of the effect of different measures on the future spread of the virus by administrators and health officials.

This paper uses both transmission rate and meteorological parameters (temperature and humidity) as features for training a set of Recurrent Neural Networks (RNN) to forecast the number of future cases of COVID-19. A systematic procedure is proposed in this paper which decomposes each signal (all features as well as the signal to be predicted) into its stationary and non-stationary modes. All the stationary modes that are similar in center frequency are used to train a separate RNN. Similarly, all the non-stationary modes are used to train another RNN. The results of all of the predictions are summed as the final forecast number of COVID-19 cases.

India is one of the highly impacted countries, and has been severely hit by the spread of the COVID-19 virus in many of its states. Some researchers have sought to predict the effect of lockdown measures on the spread of the virus in India and have suggested some policies to be followed by the jurisdictions to fight further spread of the virus in the country [16]–[18]. In [16], [17], an evolutionary data analytical method called genetic programming was used to predict the possible impact of COVID-19 in India. Here only two parameters, namely confirmed cases and total death count, were taken into consideration to analyze and predict the total rise in the coming ten days. The present work extends the former basic parametric analysis, adding transmission rate from outbreak and the local meteorological temperature and humidity data. In this paper, the data from the outbreak in different parts of India have been taken as the case study. The source of dataset for comparison is available at [19].

II. Calculation of the Transmission Rate

The number of the confirmed cases has continually increased since 24 March 2020 when an outbreak was declared in different states of India. Figure 1 shows the number of daily new confirmed cases in two of the severely affected states, namely Maharashtra (Figure 1(a)) and Tamil Nadu (Figure 1(b)), since 24 of March. These two states are taken as the case studies in this paper.

There have been overall, five lockdown periods in India as of 24 March, 2020, followed by an unlock phase. The information about each lockdown phase is outlined in the Table I. Using the number of confirmed cases corresponding to each lockdown phase, the value of the transmission rate in that phase has been calculated through the following formula [20],

$$\beta = -\frac{1}{T}\log\left(1 - I_N\left(\frac{1}{I}+\frac{1}{S}\right)\right) \tag{1}$$

where β is the mean estimated transmission rate for each lockdown phase, I_N is the number of new infections since the previous lockdown, I and S represent respectively the number of infected and susceptible individuals, and T is the sampling interval.

The transmission rates corresponding to the outbreak in Maharashtra and Tamil Nadu have been calculated using (1) for all lockdown phases of Table I. Note that the number of susceptible cases S in each state has been considered to be the entire population of that state. Figures 2(a) and 2(b) show respectively the calculated transmission rates in Maharashtra

Before constructing an RNN, we propose to pre-analyse the data to explore the nature of the signals used for training.

and Tamil Nadu per day. As can be seen from the figures, the transmission rate graphs resemble step functions. It is hypothesised that the effect of lockdown does not have immediate effects on the transmission rate as it takes time for the entire population to adapt their behaviour to the new set of rules. A robust spline based smoothing technique is exploited to slightly smooth these graphs. The so-called smoothing technique aims at balancing the fidelity in the data by minimising the following goal function [21],

$$F(\hat{S}(t)) = \| \hat{S}(t) - S(t) \|^2 + rP(\hat{S}(t)), \qquad (2)$$

where $S(t)$ and $\hat{S}(t)$ represent respectively the original and smoothed signals, and $P(\hat{S}(t))$ is a penalty term that reflects the roughness of the obtained smoothed signal ($\hat{S}(t)$). The real positive scalar parameter r is the smoothing factor that controls the degree of smoothness in $\hat{S}(t)$. A smoothing factor of $r = 10$ has been used to this end. Figures 2(c) and 2(d) show the smoothed graphs of transmission rates for Maharashtra and Tamil Nadu, respectively.

III. Signal Pre-Processing

This section presents the procedure of the proposed method, aiming to obtain a computational model that can forecast the number of confirmed cases of the COVID-19 in Maharashtra and Tamil Nadu. RNN has been widely used for time series forecasting; in this section we propose a novel framework for optimum training of an RNN using time series analysis of the signals used in training.

To train a supervised Artificial Neural Network (ANN), one needs to decide the features and labels to be used for training. Here we discuss how these features and labels can be selected systematically. As stated earlier, we hypothesise that the number of confirmed cases of COVID-19 is a function of the variability of environmental conditions (temperature and humidity), and the measures put in place by the jurisdictions to control the spread of the virus (transmission rates). The effectiveness of such measures is usually reflected by the transmission rates varying with different lockdown phases. As a result, there are four different time series introduced to the training process in this paper: (1) temperature (T), (2) humidity (H), (3) the number of confirmed cases (CC), and (4) the transmission rates (TR). Restated, this paper aims to construct an RNN to predict the future number of CC signal based on its previous observed numbers and the other aforementioned signals (T, H, and TR).

A. Time Series Analysis of Features

Before constructing an RNN, we propose to pre-analyse the data to explore the nature of the signals used for training. Since the signals used in this paper have a stochastic nature, their stationary or non-stationary behaviour is first processed in this section. This will result in more accurate training and ensure much better prediction results. First, a brief definition of the stationary and non-stationary time series is presented.

The first order autoregressive process $AR(1)$ of a signal $S(t)$ is shown as

$$s_t = \phi \, s_{t-1} + \epsilon_t \qquad (3)$$

where ϵ_t is a stationary white Gaussian noise process. Three different scenarios can occur for the above $AR(1)$ model: 1)

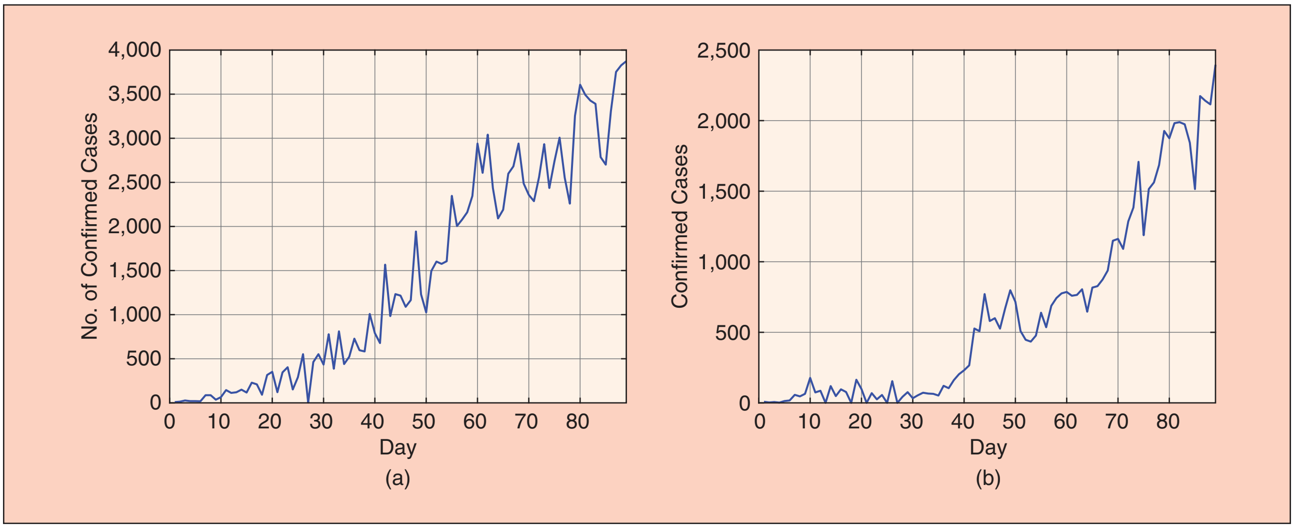


FIGURE 1 The number of daily new confirmed cases of COVID-19 in Maharashtra and Tamil Nadu as of 24 March, 2020. (a) Maharashtra. (b) Tamil Nadu.

$|\phi|<1$ implies the signal is stationary, 2) $|\phi|>1$ shows that the signal is non-stationary, and 3) $|\phi|=1$, represents a random walk model [22], [23].

In this paper, the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test is run on each signal to explore the stationary and non-stationary nature of the signal. The KPSS test is used for testing a null hypothesis of stationary time series (no unit root) around a deterministic trend (i.e. trend-stationary) against the alternative of non-stationary (unit root) [24]. The null hypothesis of trend stationary of the signal is tested against the alternative hypothesis of trend non-stationary. The test can be conducted on several autocovariance lags in the Newey-West estimator [25] of the long-run variance, each conducted at 0.1 significance level using MATLAB.

However, before running a KPSS test, one needs to select an appropriate lag length for the time series. Care must be taken to ensure that an appropriate lag length is chosen. For instance, if the lag length is too short, the test will be biased; if the lag length is too large, the power of the test will suffer. A common rule of thumb for determining the maximum lag ($L_{\max}$) can be obtained from the following equation [24],

$$L_{\max} = \left[12 \times \left(\frac{n}{100}\right)^{\frac{1}{4}}\right] \tag{4}$$

where n is the sample size and $[\cdot]$ indicates the integer part of a number. Regarding the examples of this paper $n = 80$, a maximum lag ($L_{\max}$) 11 is obtained. Three different values of 7, 9, and 11 for lags are considered for the KPSS test.

TABLE I The date of the start and end of lockdown phases in India as of 24 March, 2020.

LOCKDOWN PHASE	TIME PERIOD
I	24/03/2020–13/04/2020
II	14/04/2020–3/05/2020
III	3/05/2020–17/05/2020
IV	17/05/2020–31/05/2020
V	31/05/2020–8/06/2020
UNLOCK	8/06/2020–21/06/2020 (ONGOING)

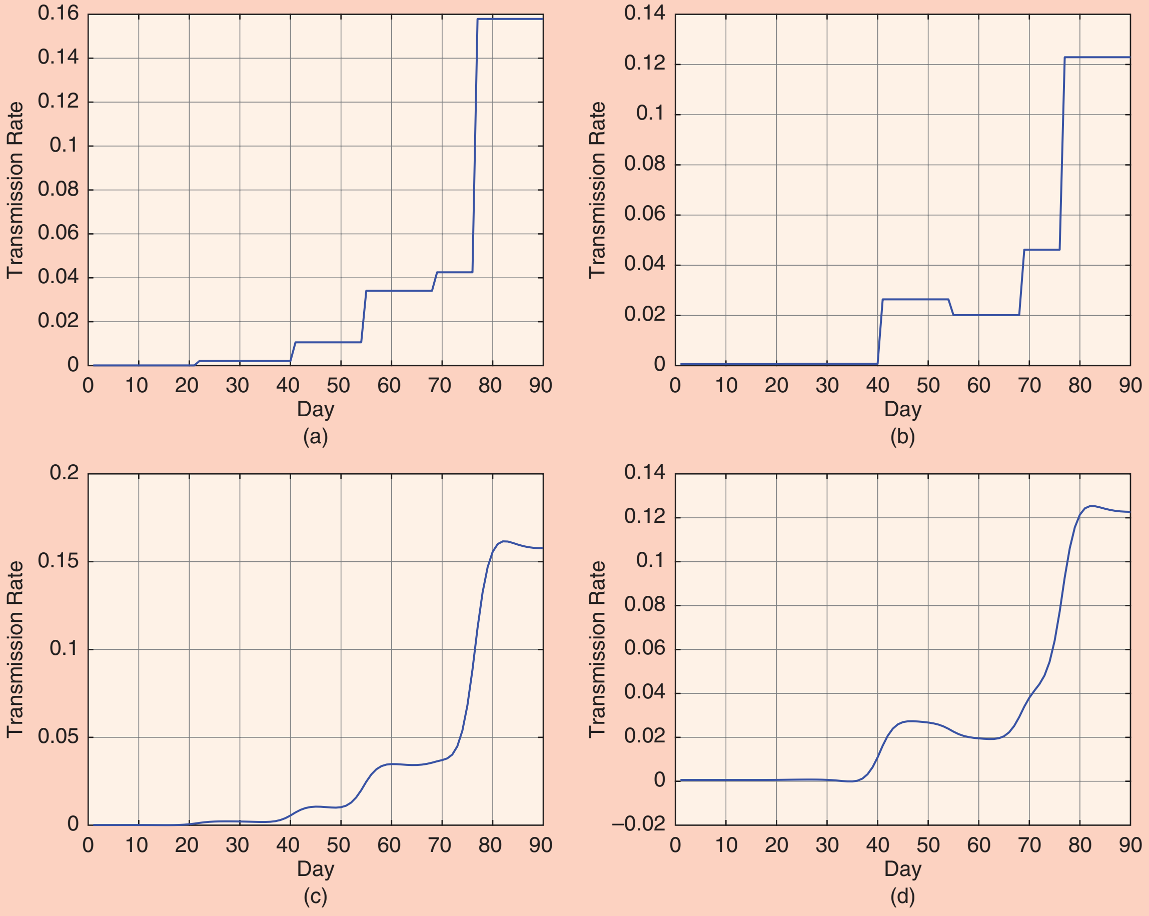


FIGURE 2 Calculated rough (a, b) and smoothed (c, d) transmission rates of COVID-19 corresponding to Maharashtra and Tamil Nadu for different lockdown phases as of 24 March, 2020. (a) Maharashtra. (b) Tamil Nadu. (c) Maharashtra, smoothed. (d) Tamil Nadu, smoothed.

Figures 3 and 4 show the temperature (T) and humidity (H) for Maharashtra and Tamil Nadu, respectively. Table II shows the results of the KPSS test run on CC, TR, T and H signals of Maharashtra.

As can be seen from the results, the KPSS test rejects the null hypothesis in favor of the alternative for the signals CC and the TR with a relatively small P-value (compared to the significance level 0.1) in all forms of the signals associated with the specified auto-covariance lags 7, 9, and 11. These signals therefore are considered non-stationary. The opposite results are obtained for the signals T and H, as can be seen from the table.

Likewise, Table III shows the results of the KPSS test run on each signal CC, TR, T, and H of Tamil Nadu. The KPSS test rejects the null hypothesis in favor of the alternative for the signals CC, TR, and H with a relatively small P-value (compared to the significance level 0.1) in all forms of the signals associated with the specified auto-covariance lags 7, 9, and 11. These signals are thus considered non-stationary. The opposite results are obtained for the signal T as seen from the table.

In the next section, more complicated signals, i.e. all signals except TR, are decomposed into some stationary and non-stationary modes using an advanced signal decomposition technique. This will further help in using features with low level of irregularities in the training process, which can further improve training results.

B. Signal Decomposition Using VMD

This section proposes to decompose complex signals (CC, T, and H) into their stationary and non-stationary oscillatory modes using an advanced decomposition technique called Variational Mode Decomposition (VMD). We further use the non-stationary part of the signals along with the signal TR

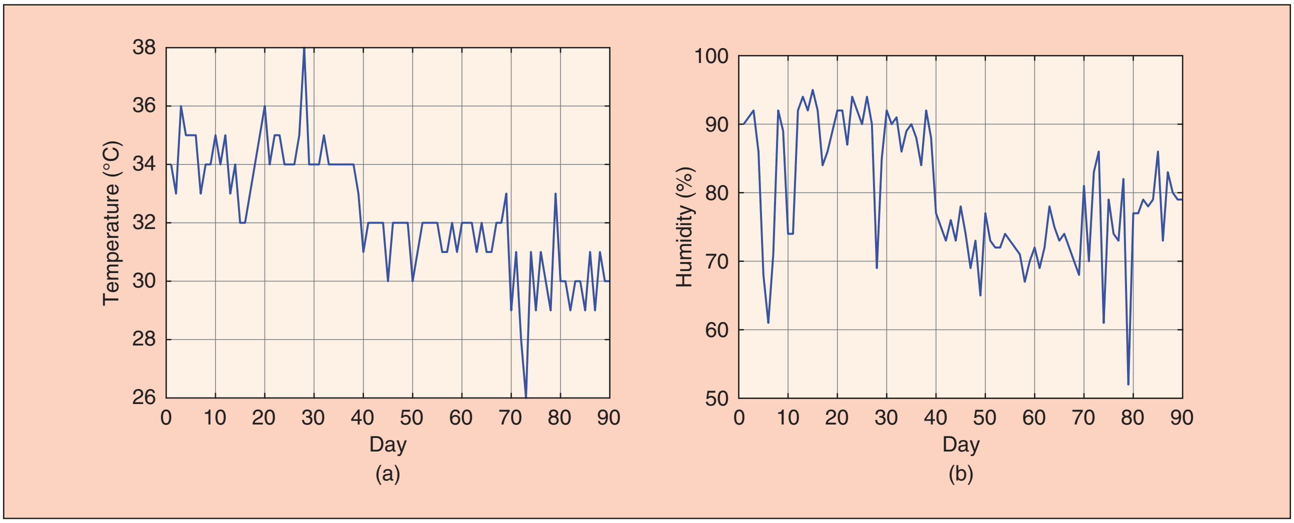


FIGURE 3 (a) Temperature, and (b) humidity time series corresponding to Maharashtra as of 24/3/2020 to 21/06/2020.

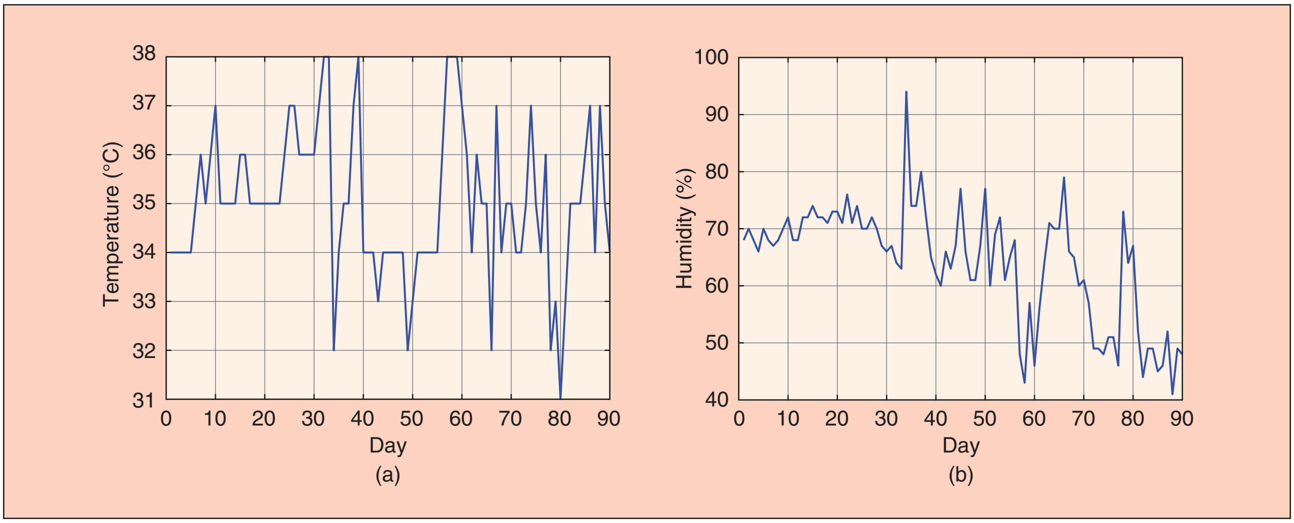


FIGURE 4 (a) Temperature, and (b) humidity time series corresponding to Tamil Nadu as of 24/3/2020 to 21/06/2020.

(non-stationary feature) for training. Similarly, the stationary modes are used to train another set of RNNs (Figure 5).

VMD is an adaptive decomposition algorithm that aims to decompose a non-linear non-stationary signal into its constructive modes [26]. These modes, which are known as Intrinsic Mode Functions (IMF), are frequency and/or amplitude modulated signals. The sum of which constructs the original signal (minus some noise, depending on settings).

VMD is an adaptive algorithm which solves a variational optimisation problem for a given signal $S(t)$ on k IMFs $\{u_k\} = \{u_1, u_2, \ldots, u_k\}$. It is assumed that each IMF is narrow-band and, therefore, has a center frequency $\{\omega_i\}$ where $i \in \{1, 2, \ldots, k\}$. The aforementioned variational optimisation problem follows,

$$\min_{\{u_k\}\&\{\omega_k\}} \sum_k \left\| \partial_t \left(\delta(t) + \frac{j}{\pi t} * u_k(t) \right) e^{-j\omega_k t} \right\|^2 \qquad (5)$$

where in the above equation, $*$ is the convolution operator. The proposers of VMD argue that the solution to the minimization problem of (5) is the saddle point of the augmented Lagrangian in a sequence of iterative sub-optimizations called alternate direction method of multipliers (ADMM) [26]. The readers are referred to the original paper for further details.

There are some critical parameters that need to be determined when using VMD for signal decomposition:

1) The number of modes (k) into which the signal is chosen to be decomposed.
2) The weight of the quadratic penalty term α, which is a denoising factor, a larger value of which admits less noise into the decomposition process. Note that in this paper, α is set to a relatively small value of 10 since denoising is not a concern [27].

TABLE II KPSS test results run on the signals corresponding to Maharashtra. Note that ST stands for stationary.

SIGNAL	LAG	P-VALUE	H	ST
CC	7, 9, 11	0.02, 0.03, 0.04	1, 1, 1	×
TR	7, 9, 11	0.01, 0.01, 0.02	1, 1, 1	×
T	7, 9, 11	0.10, 0.10, 0.10	0, 0, 0	✓
H	7, 9, 11	0.07 0.10 0.10	0, 0, 0	✓

TABLE III KPSS test results run on the signals corresponding to Tamil Nadu.

SIGNAL	LAG	P-VALUE	H	ST
CC	7, 9, 11	0.01, 0.01, 0.02	1, 1, 1	×
TR	7, 9, 11	0.01, 0.02, 0.03	1, 1, 1	×
T	7, 9, 11	0.10, 0.10, 0.10	0, 0, 0	✓
H	7, 9, 11	0.05, 0.03, 0.02	1, 1, 1	×

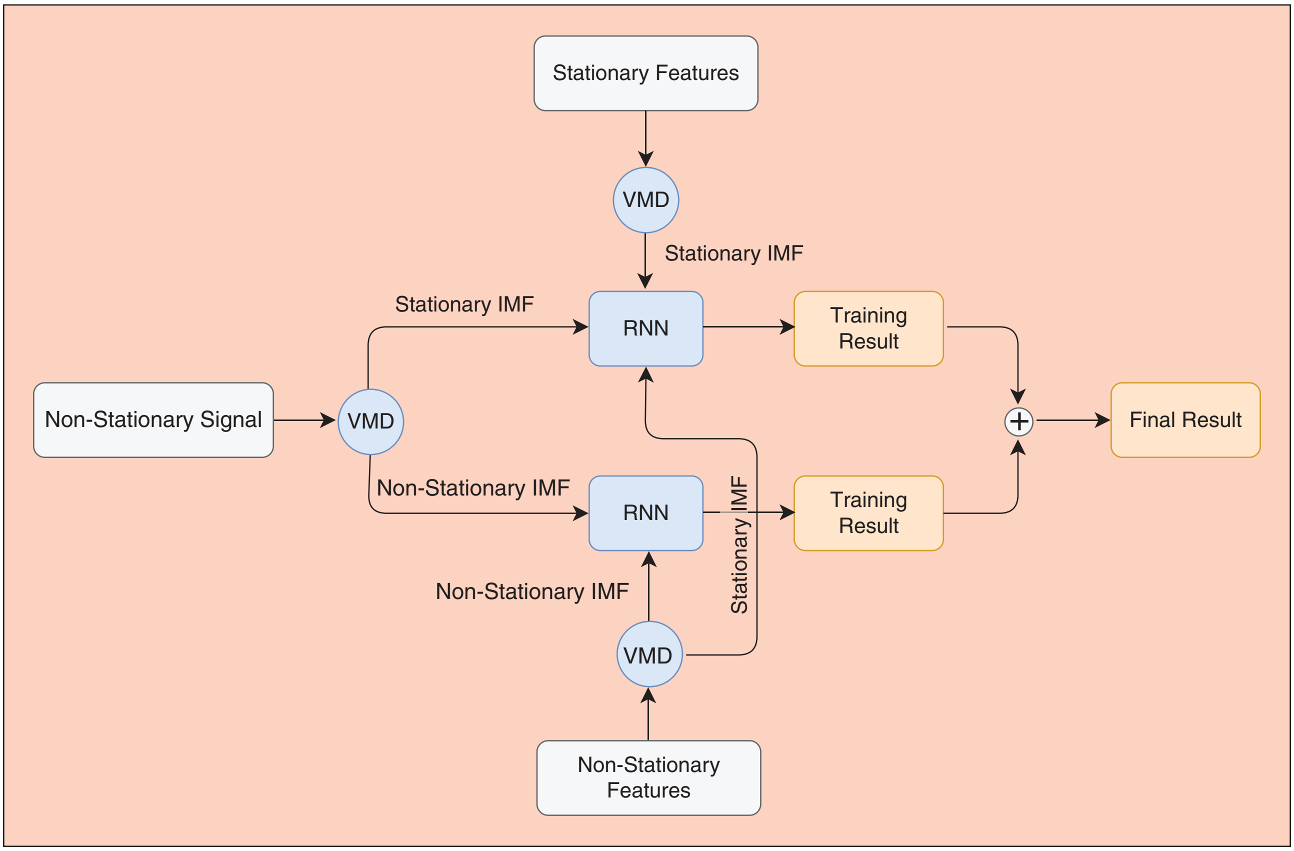


FIGURE 5 Flowchart of the proposed methodology.

3) The tolerance parameter ϵ, which controls the convergence of the algorithm. This value is set to 10^{-7} in this paper.

All signals have been decomposed into three modes. Figures 6 and 7 show respectively the IMFs corresponding to the CC signals of Maharashtra and Tamil Nadu along with their corresponding center frequencies.

We further run the KPSS test on IMFs corresponding to the VMD decomposition of the CC signals for both states using the same lags used in Section III-A, i.e. 7, 9, and 11. As is

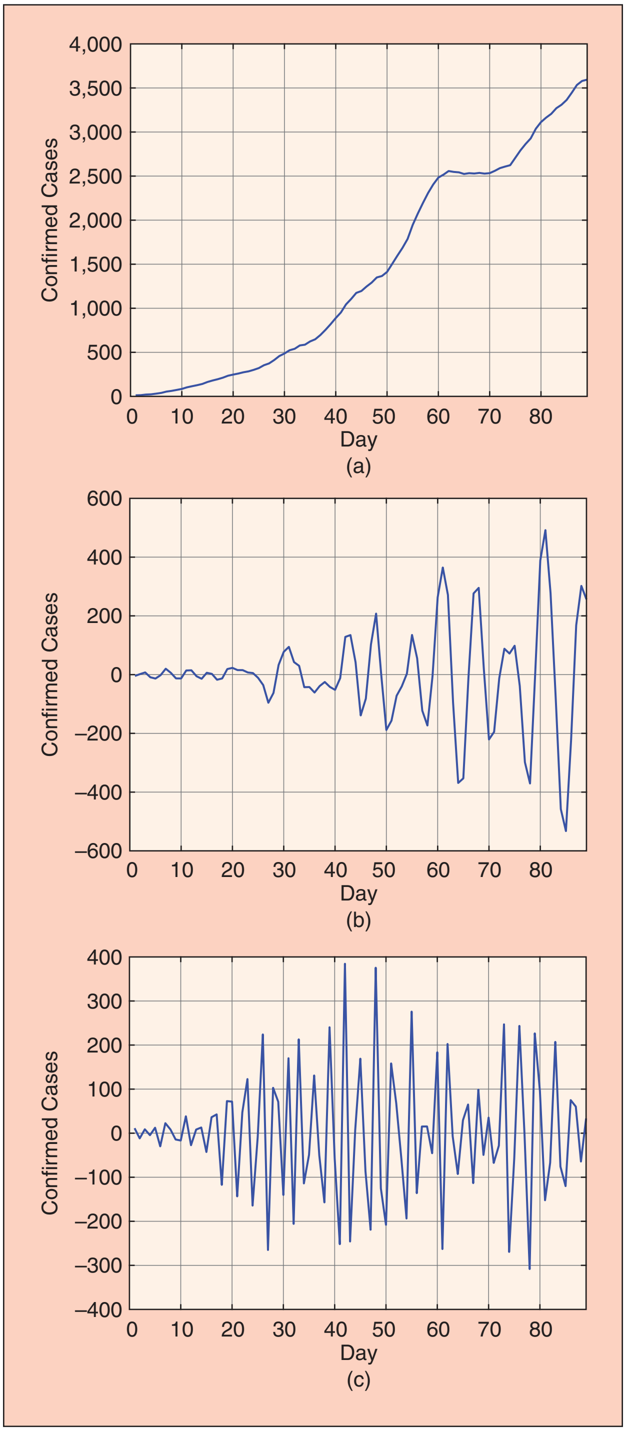


FIGURE 6 IMFs corresponding to the decomposition of CC time series of Maharashtra along with their center frequencies, (a) CC-IMF$_1$, $\omega_1 = 0.0015$. (b) CC-IMF$_2$, $\omega_2 = 0.1423$. (c) CC-IMF$_3$, $\omega_3 = 0.3480$.

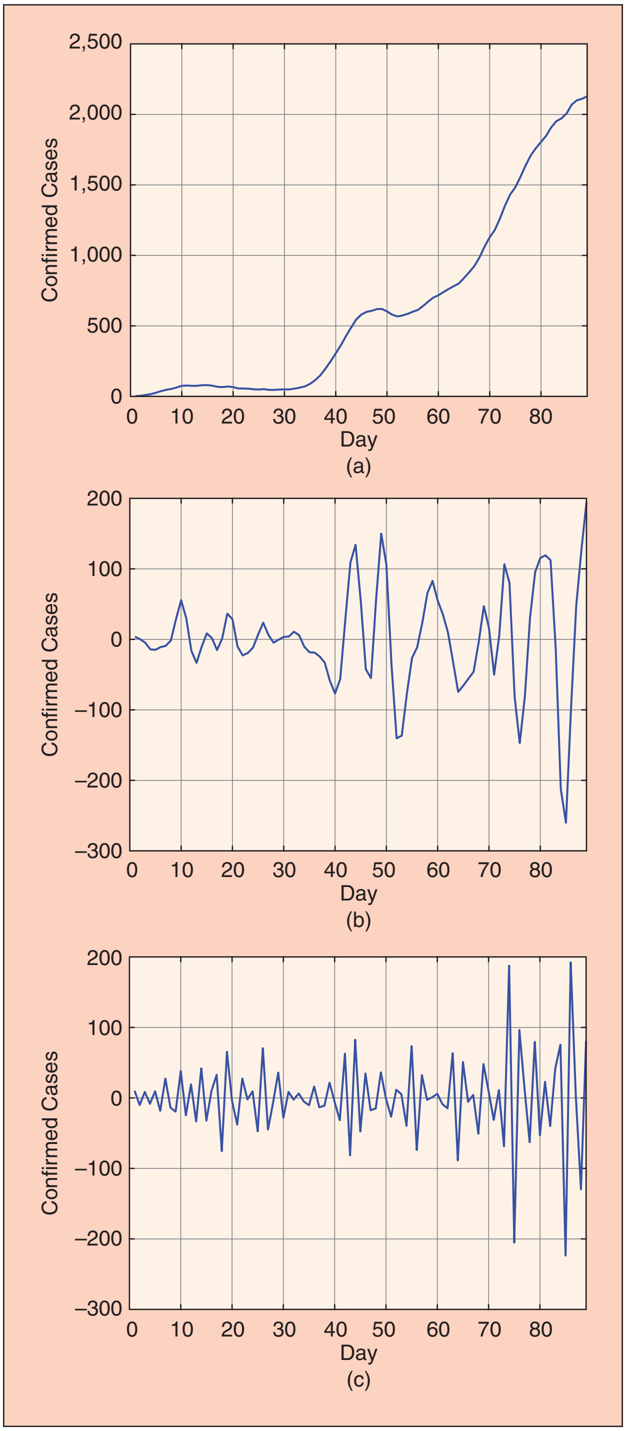


FIGURE 7 IMFs corresponding to the decomposition of CC time series of Tamil Nadu along with their center frequencies, (a) CC-IMF$_1$, $\omega_1 = 0.0025$. (b) CC-IMF$_2$, $\omega_2 = 0.1264$. (c) CC-IMF$_3$, $\omega_3 = 0.3866$.

evident from the KPSS test results, the first IMF of this decomposition in both cases is non-stationary while the remainder are stationary (Tables IV and V).

A set of multivariate stacked RNNs is developed to forecast the future values of each CC-IMF signals using the results of the previous section.

The same procedure is followed for signals T and H corresponding to Maharashtra (Figures 8 and 9) and Tamil Nadu (Figures 10 and 11). These signals are first decomposed into three IMFs, then the KPSS test is run on each IMF.

Tables VI and VII show respectively the results of KPSS tests run on IMFs of signals T and H corresponding to Maharashtra. There is no non-stationary trend in the IMFs of these signals. Likewise, Tables VIII and IX show respectively the results of KPSS tests run on IMFs of signals T and H corresponding to Tamil Nadu. As expected, regarding the H signal, H-IMF$_1$ has a non-stationary trend whereas other IMFs are stationary. As for the T signal, all IMFs show stationary trends again as expected.

The following conclusions can be made from the decomposition and KPSS test results. Regarding the decomposition of signals corresponding to Maharashtra:

1) Signals CC-IMF$_1$ (Figure 6(a)) and TR (Figure 2(c)) are trend non-stationary and, therefore, are used to train a separate RNN.
2) Signals CC-IMF$_2$ (Figure 6(b)), T-IMF$_2$ (Figure 8(b)) and H-IMF$_2$ (Figure 9(b)) are stationary and have similar center frequencies of 0.1423, 0.1507, and 0.1858 respectively and, therefore, are used to train a separate RNN.
3) Signals CC-IMF$_3$ (Figure 6(c)), T-IMF$_3$ (Figure 8(c)) and H-IMF$_3$ (Figure 9(c)) are stationary and have similar center frequencies of 0.3480, 0.3986, and 0.3860 respectively and, therefore, are used to train a separate RNN.
4) Signals T-IMF$_1$ (Figure 8(a)) and H-IMF$_1$ (Figure 9(a)) are stationary but are excluded from training process as they have no similarity to any other stationary IMFs in terms of center frequency.

Regarding decomposition of signals corresponding to Tamil Nadu:

TABLE IV KPSS test for the IMFs corresponding to the signal CC of Maharashtra.

SIGNAL	LAG	P-VALUE	H	ST
CC-IMF$_1$	7, 9, 11	0.02, 0.03, 0.04	1, 1, 1	×
CC-IMF$_2$	7, 9, 11	0.10, 0.10, 0.10	0, 0, 0	✓
CC-IMF$_3$	7, 9, 11	0.10, 0.10, 0.10	0, 0, 0	✓

TABLE V KPSS test for the IMFs corresponding to the signal CC of Tamil Nadu.

SIGNAL	LAG	P-VALUE	H	ST
CC-IMF$_1$	7, 9, 11	0.01, 0.01, 0.02	1, 1, 1	×
CC-IMF$_2$	7, 9, 11	0.10, 0.10, 0.10	0, 0, 0	✓
CC-IMF$_3$	7, 9, 11	0.10, 0.10, 0.10	0, 0, 0	✓

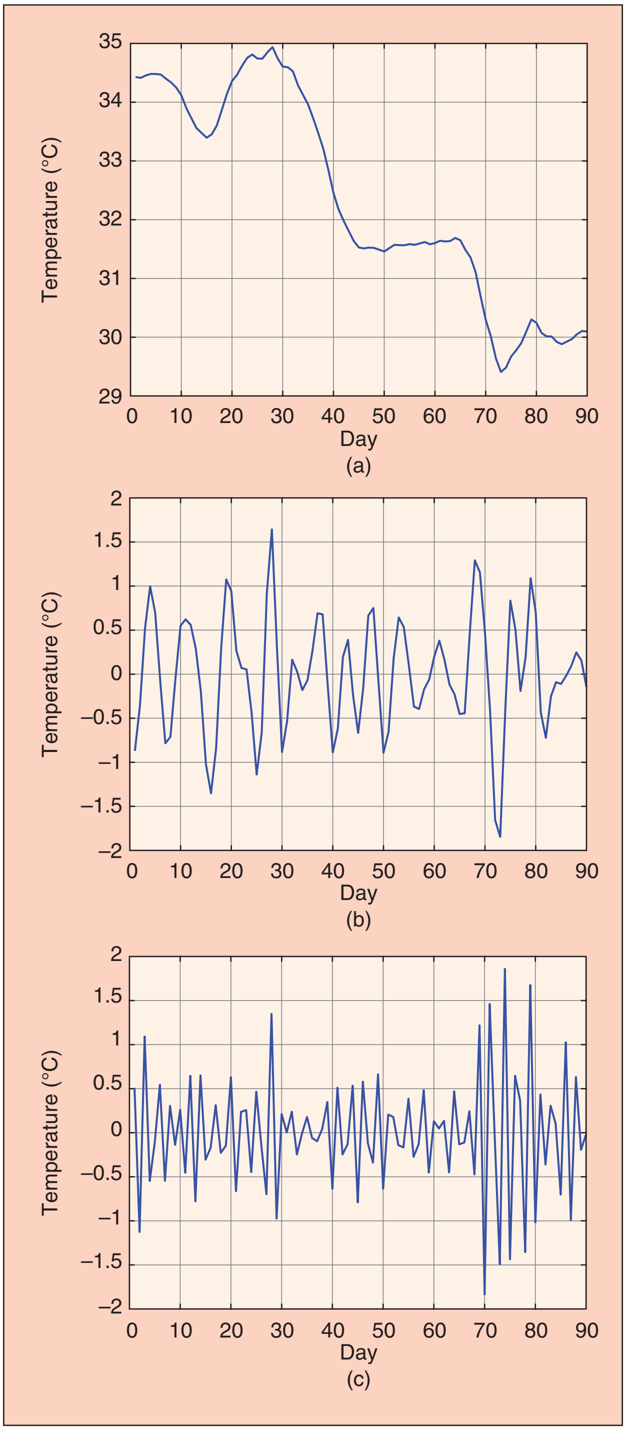


FIGURE 8 IMFs corresponding to the decomposition of temperature (T) time series of Maharashtra along with their center frequencies, (a) T-IMF$_1$, $\omega_1 = 10^{-5}$. (b) T-IMF$_2$, $\omega_2 = 0.1507$. (c) T-IMF$_3$, $\omega_3 = 0.3986$.

1) Signals CC-IMF$_1$ (Figure 7(a)), TR (Figure 2(d)), and H-IMF$_1$ (Figure 11(a)) are trend non-stationary and, therefore, are used to train a separate RNN.
2) Signals CC-IMF$_2$ (Figure 7(b)), T-IMF$_2$ (Figure 10(b)) and H-IMF$_2$ (Figure 11(b)) are stationary and have relatively similar center frequencies of 0.1264, 0.1312, and 0.0887 respectively and, therefore, are used to train a separate RNN.
3) Signals CC-IMF$_3$ (Figure 7(c)), T-IMF$_3$ (Figure 10(c)) and H-IMF$_3$ (Figure 11(c)) are stationary and have similar

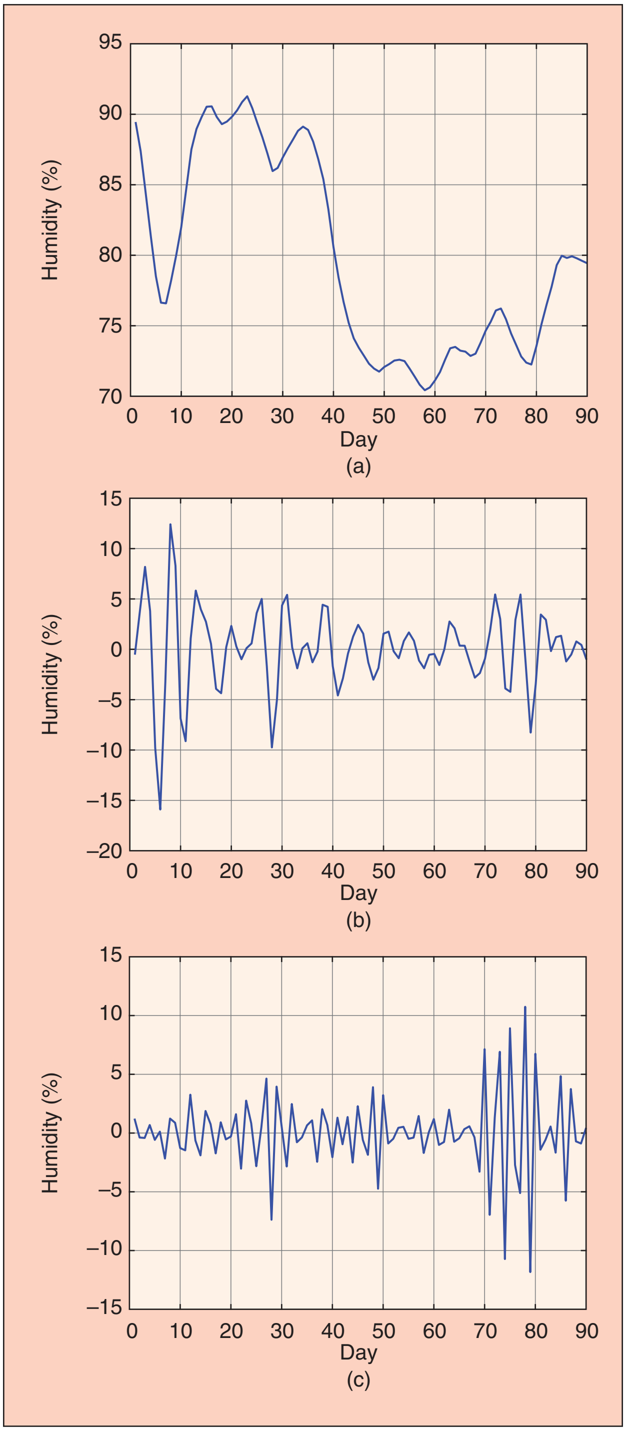


FIGURE 9 IMFs corresponding to the decomposition of humidity (H) time series of Maharashtra along with their center frequencies, (a) H-IMF$_1$, $\omega_1 = 0.0001$. (b) H-IMF$_2$, $\omega_2 = 0.1858$. (c) H-IMF$_3$, $\omega_3 = 0.3860$.

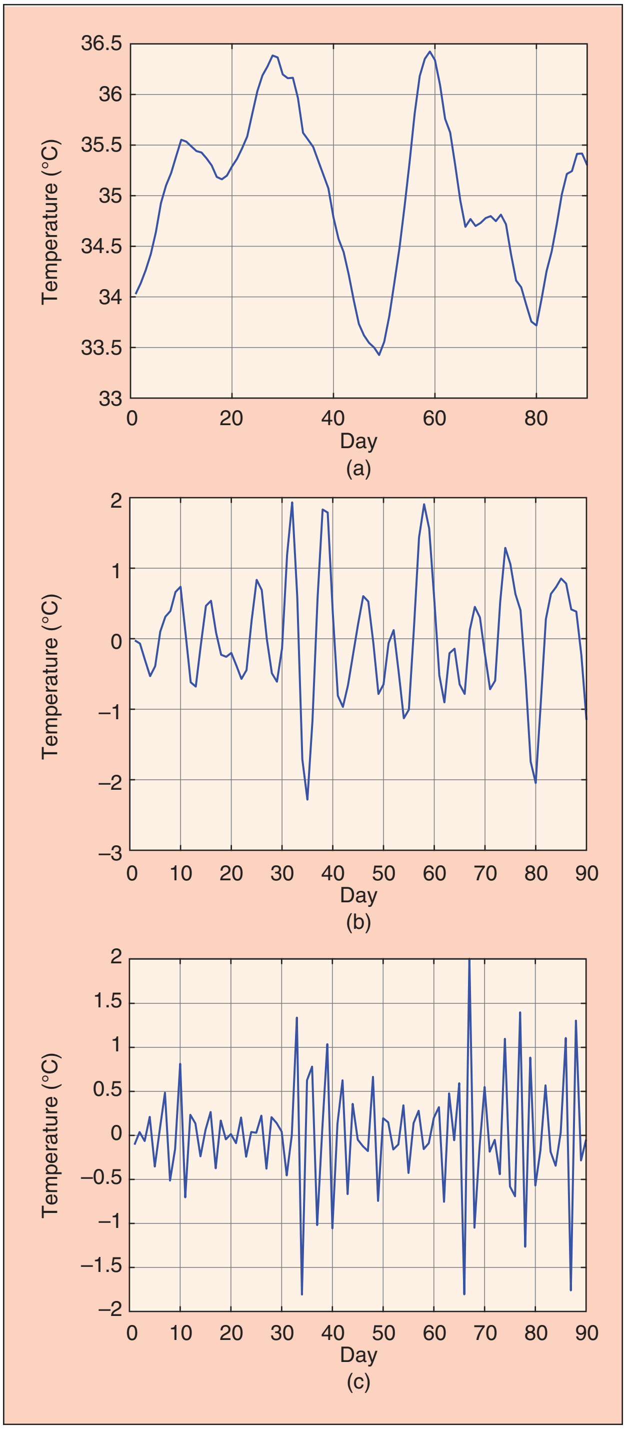


FIGURE 10 IMFs corresponding to the decomposition of temperature (T) time series of Tamil Nadu along with their center frequencies, (a) T-IMF$_1$, $\omega_1 = 9 \times 10^{-6}$. (b) T-IMF$_2$, $\omega_2 = 0.1312$. (c) T-IMF$_3$, $\omega_3 = 0.3804$.

center frequencies of 0.3866, 0.3804, and 0.3525 respectively and, therefore, are used to train a separate RNN.

4) Signal T-IMF$_1$ (Figure 10(a)) is stationary but is excluded from training process as it has no similarity to any other stationary IMFs in terms of center frequency.[1]

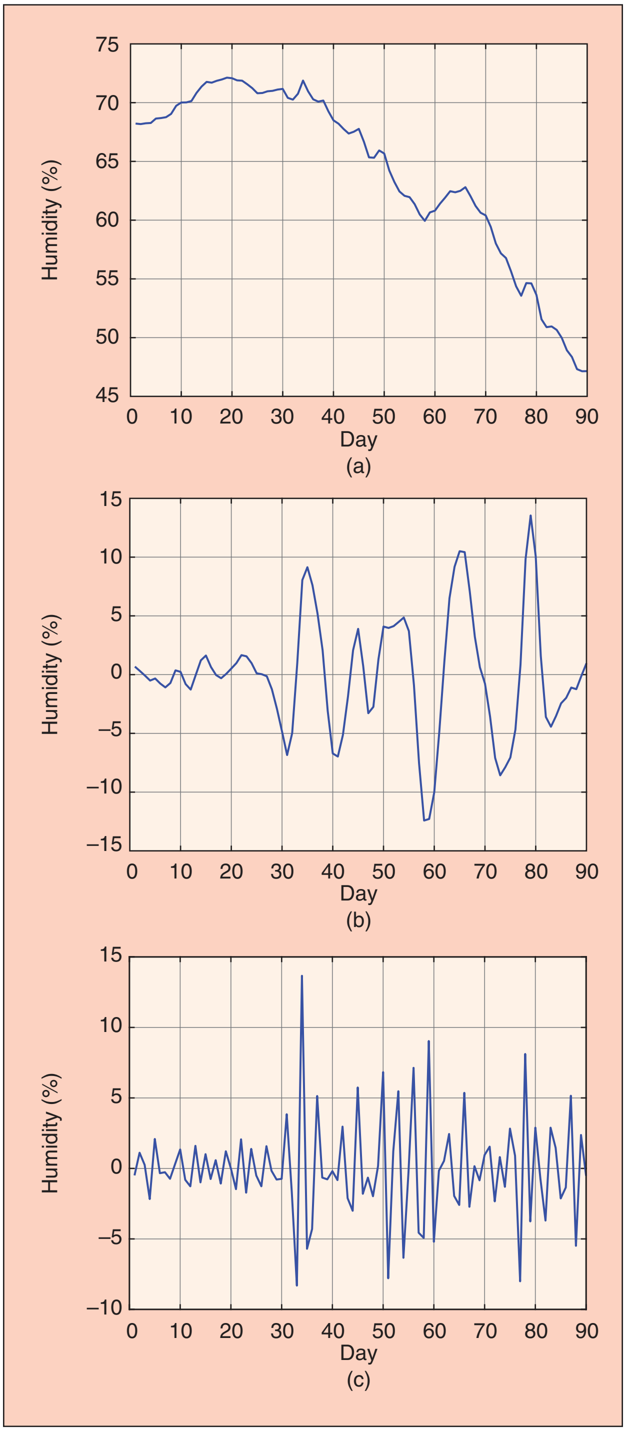


FIGURE 11 IMFs corresponding to the decomposition of humidity (H) time series of Tamil Nadu along with their center frequencies, (a) H-IMF$_1$, $\omega_1 = 0.0001$. (b) H-IMF$_2$, $\omega_2 = 0.0887$. (c) H-IMF$_3$, $\omega_3 = 0.3525$.

IV. Training Sequence Models

A set of multivariate stacked RNNs is developed to forecast the future values of each CC-IMF signals using the results of the previous section. The results of all predicted values of CC-IMFs are summed to obtain the forecast value of the CC signal one step forward in the future (Figure 5). A set of Recurrent Neural Networks (RNNs) with Long Short Term Memory (LSTM) cells is used because RNNs have been proven to be effective for forecasting time series [28], [29].

LSTM cells were initially designed to deal with vanishing and exploding gradient problems in sequence models [30]. The structure of LSTM is briefly explained in the next section.

TABLE VI KPSS test for the IMFs corresponding to the signal T of Maharashtra.

SIGNAL	LAG	P-VALUE	H	ST
T-IMF$_1$	7, 9, 11	0.10, 0.10, 0.10	0, 0, 0	✓
T-IMF$_2$	7, 9, 11	0.10, 0.10, 0.10	0, 0, 0	✓
T-IMF$_3$	7, 9, 11	0.10, 0.10, 0.10	0, 0, 0	✓

TABLE VII KPSS test for the IMFs corresponding to the signal H of Maharashtra.

SIGNAL	LAG	P-VALUE	H	ST
H-IMF$_1$	7, 9, 11	0.05, 0.09, 0.10	0, 0, 0	✓
H-IMF$_2$	7, 9, 11	0.10, 0.10, 0.10	0, 0, 0	✓
H-IMF$_3$	7, 9, 11	0.10, 0.10, 0.10	0, 0, 0	✓

TABLE VIII KPSS test for the IMFs corresponding to the signal T of Tamil Nadu.

SIGNAL	LAG	P-VALUE	H	ST
T-IMF$_1$	7, 9, 11	0.10, 0.10, 0.10	0, 0, 0	✓
T-IMF$_2$	7, 9, 11	0.10, 0.10, 0.10	0, 0, 0	✓
T-IMF$_3$	7, 9, 11	0.10, 0.10, 0.10	0, 0, 0	✓

TABLE IX KPSS test for the IMFs corresponding to the signal H of Tamil Nadu.

SIGNAL	LAG	P-VALUE	H	ST
H-IMF$_1$	7, 9, 11	0.01, 0.01, 0.02	1, 1, 1	×
H-IMF$_2$	7, 9, 11	0.10, 0.10, 0.10	0, 0, 0	✓
H-IMF$_3$	7, 9, 11	0.10, 0.10, 0.10	0, 0, 0	✓

[1] Note that one may argue that the first IMF of the temperature and humidity signals in any cases has a non-stationary trend in the long run and, therefore, suggest to consider them as features for training CC-IMF$_1$. The authors decided to ignore them to avoid masking the effect of non-stationary features which are believed to have more impact on CC-IMF$_1$.

A. Long Short Term Memory (LSTM) Cells

An LSTM unit consists of three gates (i.e., update, forget, and output gates) and three cells (i.e., input, memory, and update cells). The memory cell at time t is updated using a candidate value $\tilde{c}^{<t>}$, which is calculated using the output value at time $t-1$, i.e., $a^{<t-1>}$, and input value at time t, i.e., $x^{<t>}$, through the equation

$$\tilde{c}^{<t>} = \tanh(W_c[a^{<t-1>}, x^{<t>}] + b_c) \tag{6}$$

where $\tanh(.)$ is the hyperbolic tangent activation function, and W_c and b_c represent the matrix of parameters and biased vector of the memory cell, respectively. The value of the memory cell $c^{<t>}$ is then updated using the candidate value $\tilde{c}^{<t>}$ and the previous value $c^{<t-1>}$ through

$$c^{<t>} = \Gamma_u \odot \tilde{c}^{<t>} + \Gamma_f \odot c^{<t-1>} \tag{7}$$

where $\odot$ indicates element-wise multiplication. Γ_u and Γ_f are the values of the update and forget gates which are obtained from

$$\Gamma_u = \sigma(W_u[a^{<t-1>}, x^{<t>}] + b_u) \tag{8}$$

and

$$\Gamma_f = \sigma(W_f[a^{<t-1>}, x^{<t>}] + b_f) \tag{9}$$

in which $\sigma(.)$ is the sigmoid activation function, W_u and b_u are respectively the matrix of parameters and the bias vector corresponding to the update gate, and W_f and b_f are respectively the matrix of parameters and the bias vector corresponding to the forget gate.

The output value of the LSTM unit at time t is

$$a^{<t>} = \Gamma_o \odot \tanh(c^{<t>}) \tag{10}$$

where Γ_o is the value of the output gate which itself is

$$\Gamma_o = \sigma(W_o[a^{<t-1>}, x^{<t>}] + b_o) \tag{11}$$

in which W_o and b_o are respectively the matrix of parameters and the bias vector corresponding to the output gate. Figure 12 shows an LSTM unit.

A multivariate RNN architecture is used in this paper, which takes multiple features as input, and outputs the predicted value. Two different architectures are used, one for training CC-IMF$_1$, and another for training CC-IMF$_2$ and CC-IMF$_3$ separately. The architecture of the stacked RNN corresponding to the forecasting future value of CC-IMF$_1$ is as follows:

1) a sequence input layer which accepts the number of inputs equal to the number of features. Regarding Maharashtra, there are two features for training CC-IMF$_1$: the signals CC-IMF$_1$ and TR at time $t-1$. For Tamil Nadu there are three features for training CC-IMF$_1$: the signals CC-IMF$_1$, H-IMF$_1$ and TR at time $t-1$. The value of the CC-IMF$_1$ at time t is the target value which needs to be predicted in both cases.
2) an LSTM layer with 50 units.
3) a dropout layer with the factor 0.6.
4) a fully connected layer with one output unit.

The architecture used for training CC-IMF$_2$ and CC-IMF$_3$ is as follows:

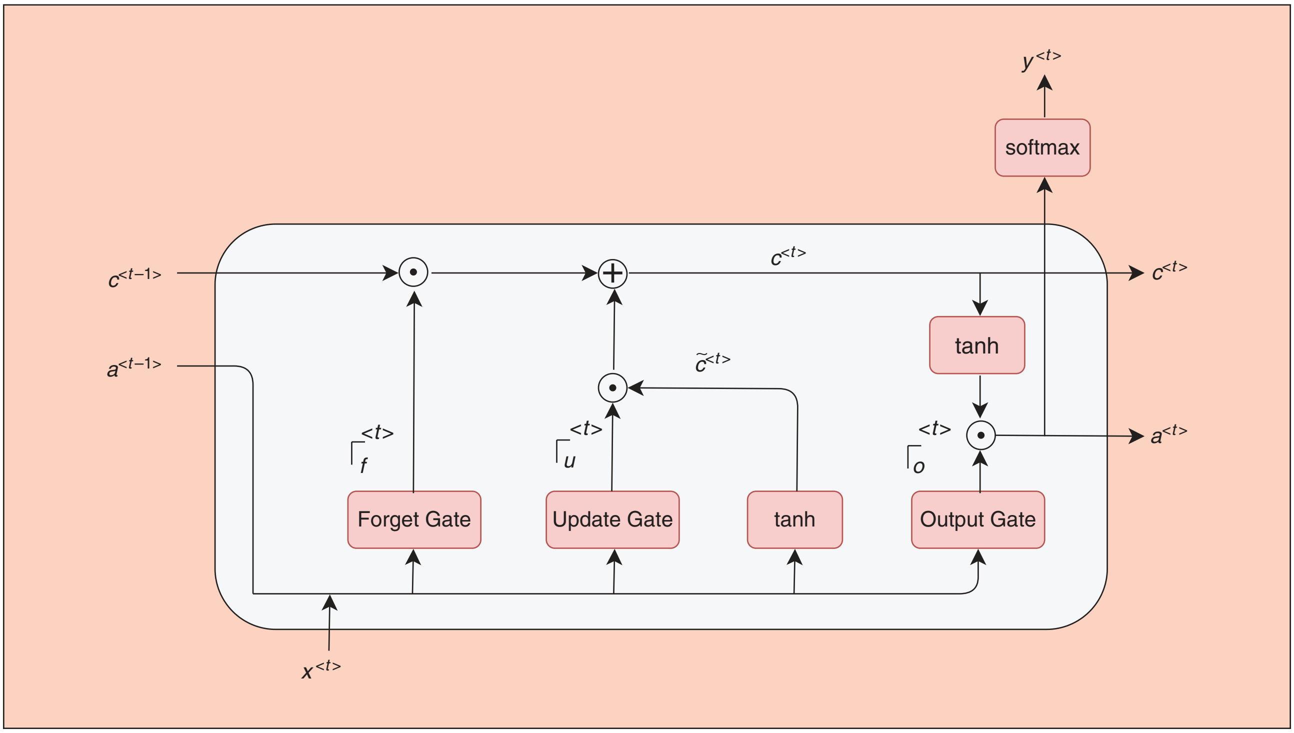


FIGURE 12 Visualisation of an LSTM unit. $y^{<t>}$ is the final output of an LSTM unit at time t which is computed by a softmax activation function.

1) a sequence input layer which accepts the number of inputs equal to the number of features. Note that there are three features for training CC-IMF$_2$ in both cases: the signals CC-IMF$_2$, T-IMF$_2$ and H-IMF$_2$ at time $t-1$. The target for the network in this case is the value of the signal CC-IMF$_2$ at time t. Similarly, there are three features for training CC-IMF$_3$ in both cases: the signals CC-IMF$_3$, T-IMF$_3$ and H-IMF$_3$ at time $t-1$. The target for the network is then the value of CC-IMF$_3$ at time t.
2) an LSTM layer with 200 units.
3) a fully connected layer with 200 units.
4) a dropout layer with the factor 0.6.

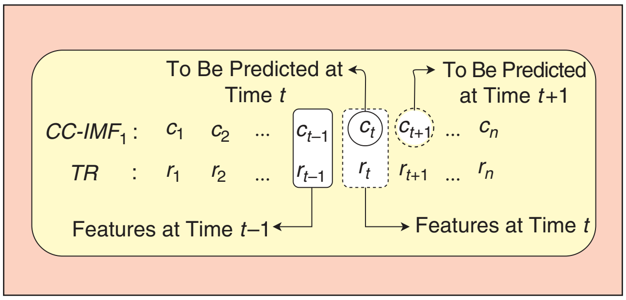

To Be Predicted at Time t
To Be Predicted at Time t+1
CC-IMF$_1$: c_1 c_2 ... c_{t-1} c_t c_{t+1} ... c_n
TR : r_1 r_2 ... r_{t-1} r_t r_{t+1} ... r_n
Features at Time t–1
Features at Time t

FIGURE 13 The value of the signal CC-IMF$_1$ at time t (c_t) is predicted using its value at time $t-1$ (c_{t-1}) and the value of the signal TR at time $t-1$ (r_{t-1}) as features.

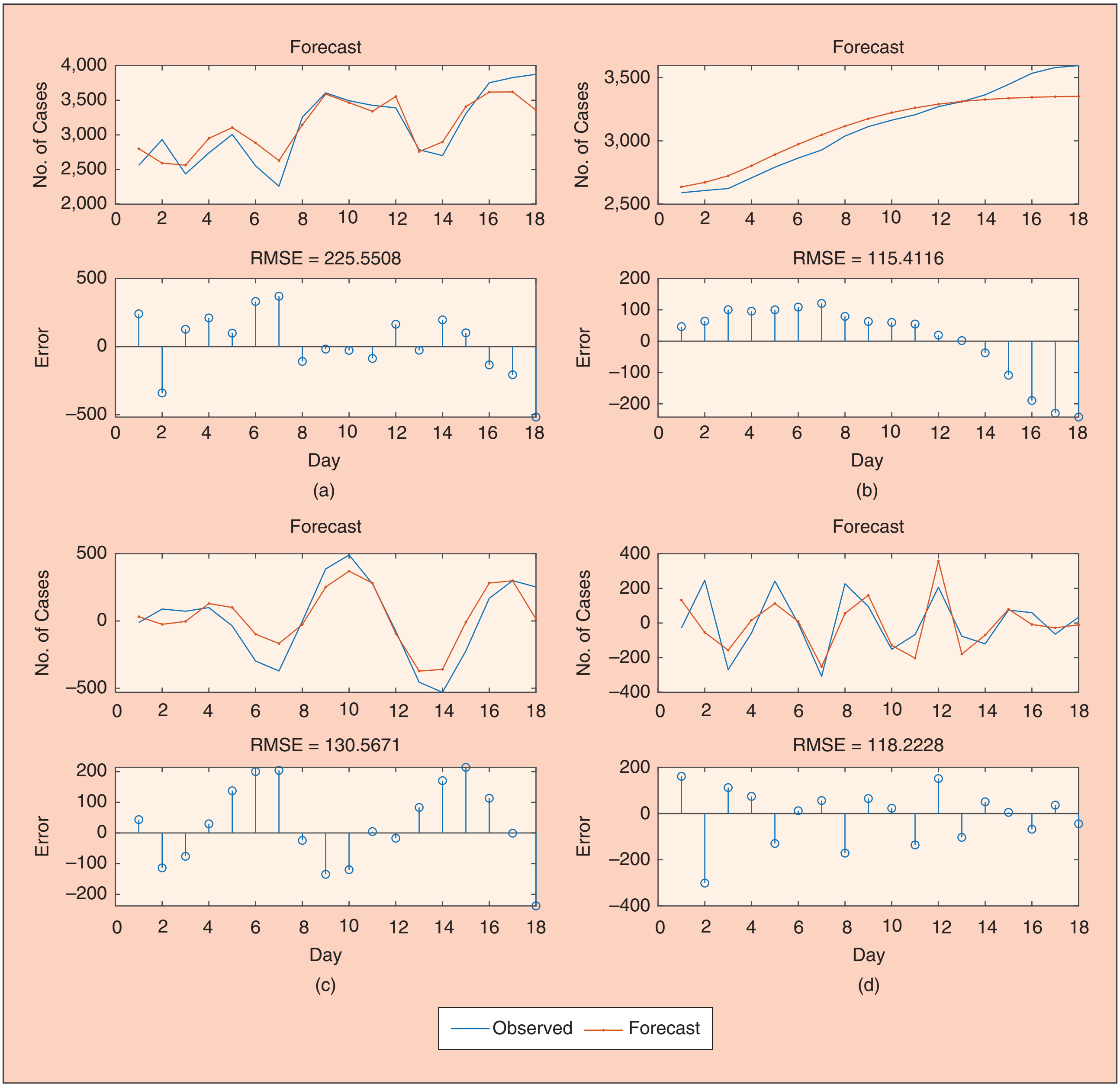


FIGURE 14 Predicted value and root mean square error (RMSE) of the number of COVID-19 cases forecast corresponding to each IMF of Maharashtra CC signal conducted on the test set. (a) CC; (b) CC-IMF$_1$; (c) CC-IMF$_2$; and (d) CC-IMF$_3$.

5) an LSTM layer with 50 units.

6) a fully connected layer with one output unit.

The dropout layers were adapted to prevent over-fitting on the training set. In fact, the dropout hyper-parameter indicates the probability of training a given node in a layer. It has the regularisation effect and prevents over-fitting on the training set [31].

B. Construction of the Training Set and the Test Set

Consider signals CC-IMF$_1$ and TR corresponding to Maharashtra. The proportion of 80% of data in both signals is used in the training set and the remainder (20%) is considered as the test set for validation purpose. To construct the training set, the value of the signal CC-IMF$_1$ and TR at time $t-1$ are considered as features to be fed into the constructed RNN, and the value of the signal CC-IMF$_1$ at time t is considered as the label or expected output of the RNN (Figure 13). The size of the training set is equal to 80% of the number of elements of signal CC-IMF$_1$, rounded to an integer (or equivalently the integer part of 80% of the number of elements of signal TR). The same procedure is followed in all other cases.

C. Setting the Options for the RNNs

Adam optimisation has been set in options as the optimisation method to update network weights in each iteration,

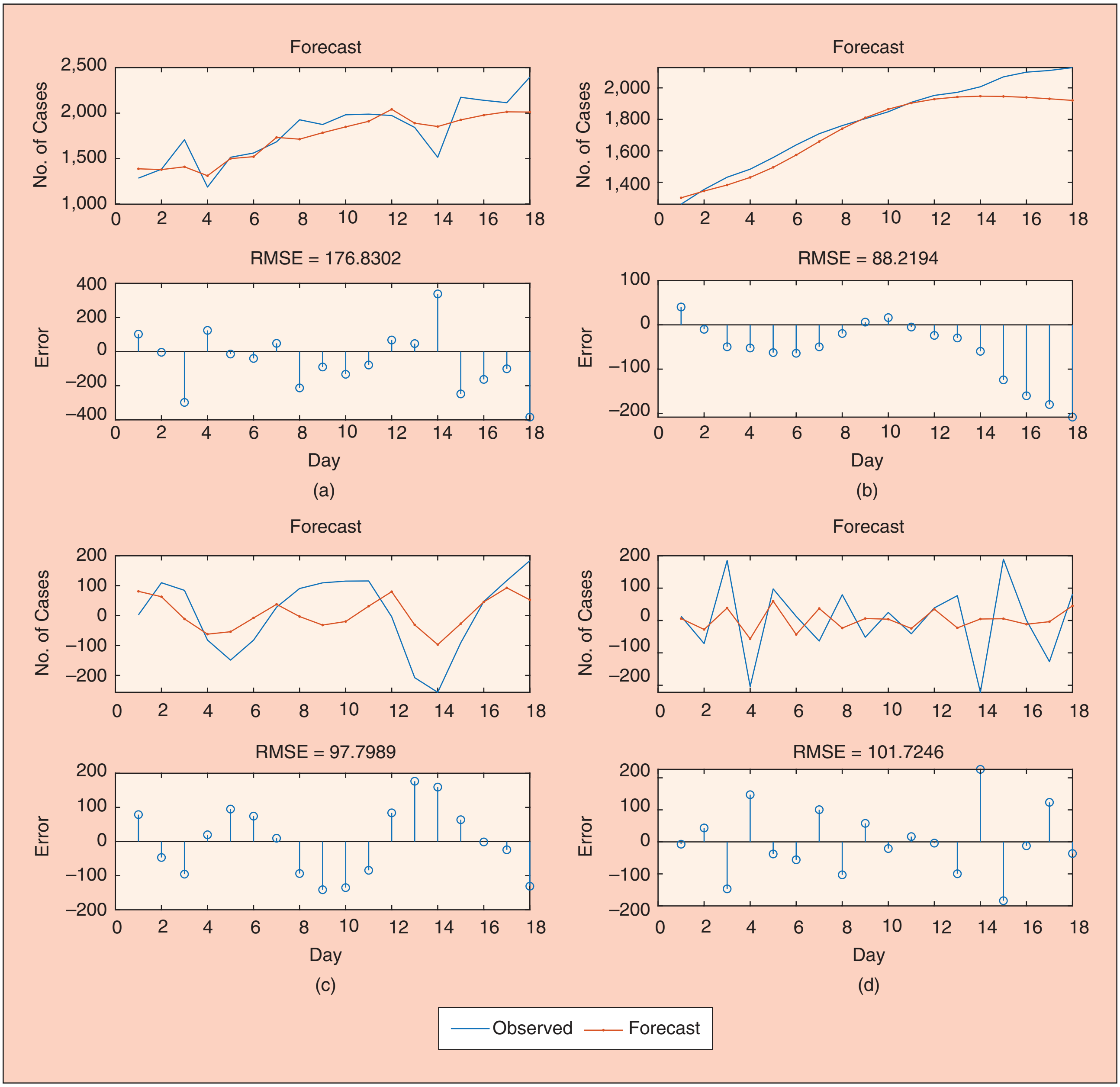


FIGURE 15 Predicted value and Root mean square error (RMSE) of the number of COVID-19 cases forecast corresponding to each IMF of Tamil Nadu CC signal conducted on the test set. (a) CC; (b) CC-IMF$_1$; (c) CC-IMF$_2$; and (d) CC-IMF$_3$.

as it is known to be an adaptive learning rate optimization algorithm designed specifically for training deep neural networks [32]. The learning rate was set initially at 0.005 and was decreased by a factor of 0.2 at every 200 epochs. The number of maximum epochs was chosen to be 1000. In order to avoid exploding gradients effect, a threshold 1 was set as the gradient threshold.

> ... the decomposed IMFs with similar center frequencies are used to train separate RNNs. Here we further work out the phase of each decomposed IMF corresponding to the CC, T, and H signals for both states using Gabor's complex analytical signal.

V. Results and Discussion

Figure 14 shows the predicted number of cases of COVID-19 for CC-IMF$_1$ (Figure 14(b)), CC-IMF$_2$ (Figure 14(c)) and CC-IMF$_3$ (Figure 14(d)) and the sum of all of them (Figure 14(a)) for the state of Maharashtra conducted on the test set. The figures show the Root Mean Square Error (RMSE) corresponding to each case. The results show that the model can predict the future number of cases within an acceptable range of error. However, Figure 14(b) shows that the predicted value of the signal deviates from its expected value at the right end of the signal. This is likely due to the end effect arising from the spline method used in the smoothing procedure to smooth the feature TR. This effect is also evident in Figure 2(c). As can be seen in the figure, the right end of the signal is slightly tilted downward whereas this is not the case in the original signal of Figure 2(a). Therefore, one may argue that smoothing the

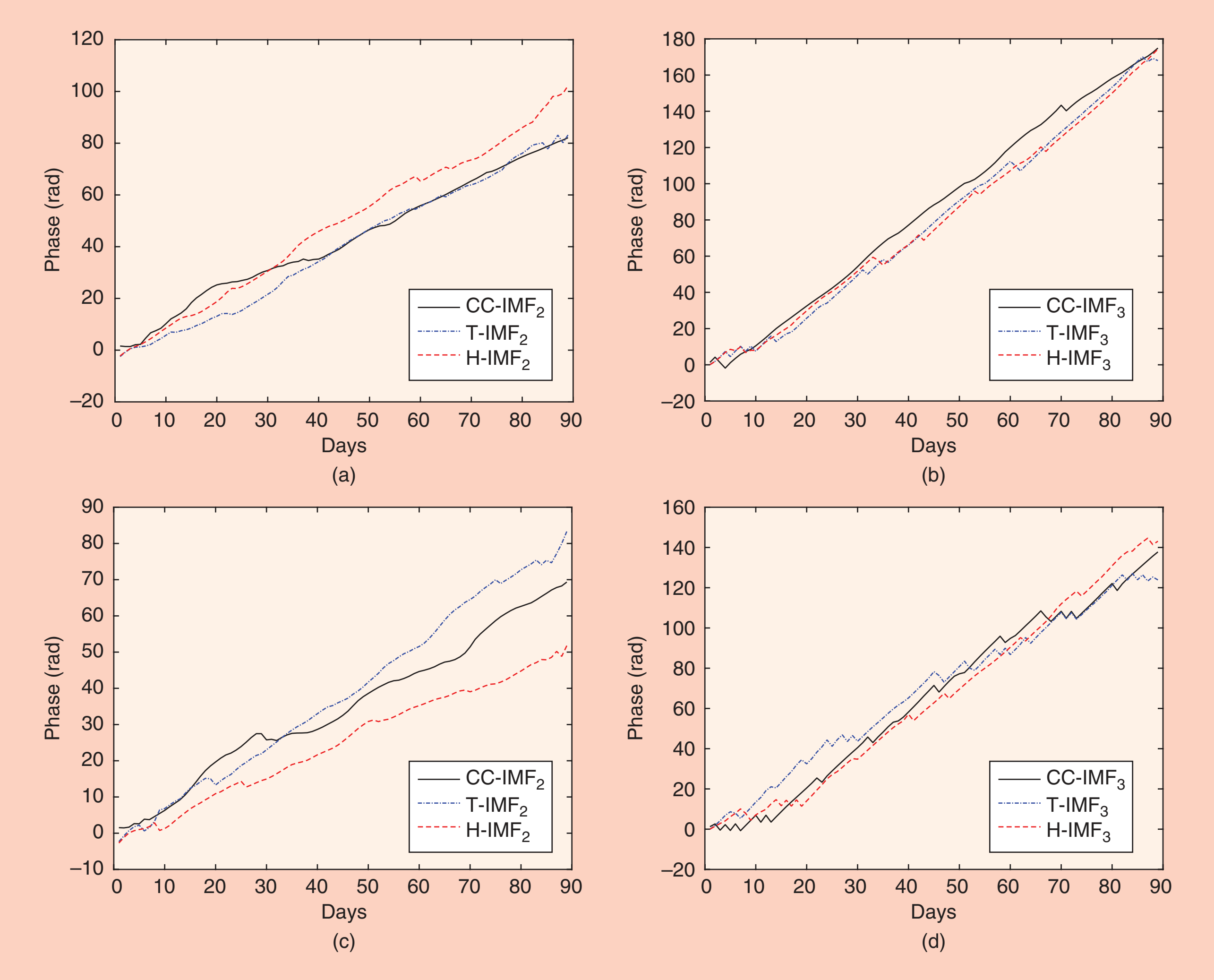


FIGURE 16 Unwraped phase corresponding to the IMFs of CC, T and H signals of states Maharashtra and Tamil Nadu. (a) Maharashtra (IMF$_2$). (b) Maharashtra (IMF$_3$). (c) Tamil Nadu (IMF$_2$). (d) Tamil Nadu (IMF$_3$).

signal TR is not beneficial. However, the prediction results have been boosted when the TR signal was smoothed.

The same procedure has been followed to train RNNs to predict the CC signal corresponding to Tamil Nadu. The results of the trained RNN on the test set is presented in Figure 15. The same effect of smoothing the TR signal is evident in Figure 15(a). The second and third modes of the CC signal are not as accurate as those of Maharashtra. The reason is that we conformed to the same architecture which was developed initially for Maharashtra. In the following we discuss this in more details.

We first look into the mean absolute percentage error (MAPE) corresponding to predictions for both cases, in order to compare the precision of the two different forecast problems with one another. The MAPE is calculated as

$$\mathrm{MAPE} = \mathrm{mean}\left(100 \times \left|\frac{p_t - y_t}{y_t}\right|\right) \tag{12}$$

where p_t and y_t represent respectively the predicted and observed values of the time series. The MAPEs for Maharashtra and Tamil Nadu are 6.23% and 7.77%, respectively.

As explained in Section III-B, the decomposed IMFs with similar center frequencies are used to train separate RNNs. Here we further work out the phase of each decomposed IMF corresponding to the CC, T, and H signals for both states using Gabor's complex analytical signal $X_a(t)$ [33] which is defined as

$$X_a(t) = X(t) + j\hat{X}(t), \tag{13}$$

where $X(t)$ and $\hat{X}(t)$ are respectively the original signal and its Hilbert transform. One can obtain the instantaneous phase of each band IMF as follows,

$$\phi(t) = \tan^{-1}\left(\frac{\hat{X}(t)}{X(t)}\right). \tag{14}$$

Figure 16 shows the obtained unwrapped phase of the IMFs corresponding to the aforementioned signals for both states (cf. `unwrap()` in Matlab). From Figures 16(a) and 16(b), the phase of the IMFs corresponding to the CC, H, and T signals of Maharashtra are more synchronised compared to those of

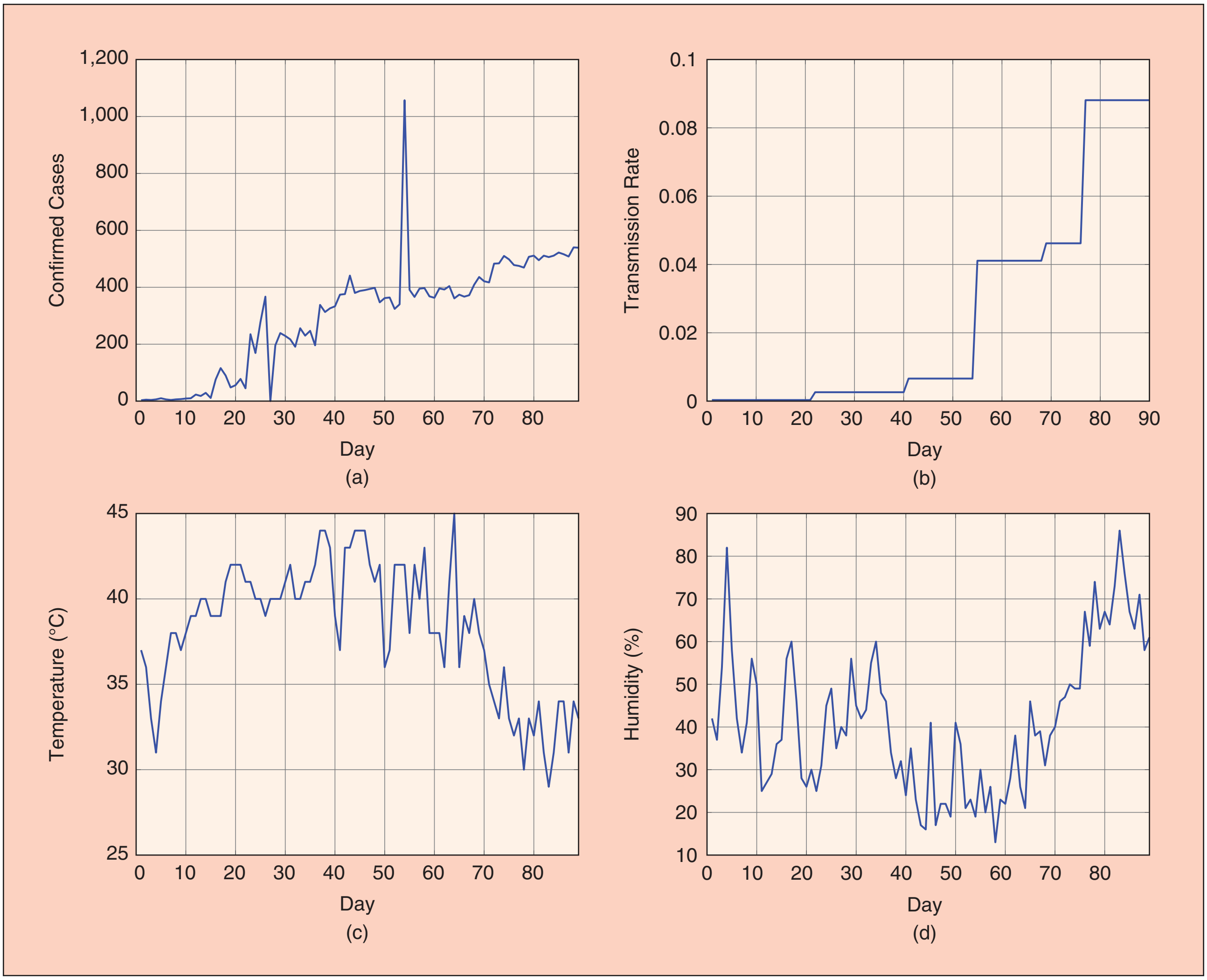


FIGURE 17 Signals (a) CC; (b) TR; (c) T; and (d) H corresponding to the state Gujarat.

Tamil Nadu (Figures 16(c) and 16(d)). This suggests a more complex dependency among IMFs of these signals corresponding to Tamil Nadu compared with Maharashtra. One way of achieving more accuracy in prediction in the case of Tamil Nadu is to use a deeper RNN architecture. However, in order to avoid over-fitting, either more data or a more severe regularisation strategy has to be exploited.

A further example is now investigated, corresponding to the state Gujarat in India where the number of cases is smaller. The data from this state is of interest particularly due to an outlier presenting at around day 54 (Figure 17(a)). The transmission rate TR of Figure 17(b) has been smoothed using the technique proposed in Section II. Also, all the signals CC, T, and H are decomposed using VMD (Section III-B), and their stationary and non-stationary parts are grouped and used for training the RNNs as discussed respectively in Sections III-B and IV. Figure 18 shows the final results of the prediction process. A satisfactory value of MAPE = 4.68% is obtained, which further confirms the applicability of the proposed technique.

VI. Conclusion

A systematic procedure to derive features for training RNNs to forecast the future number of confirmed cases of COVID-19

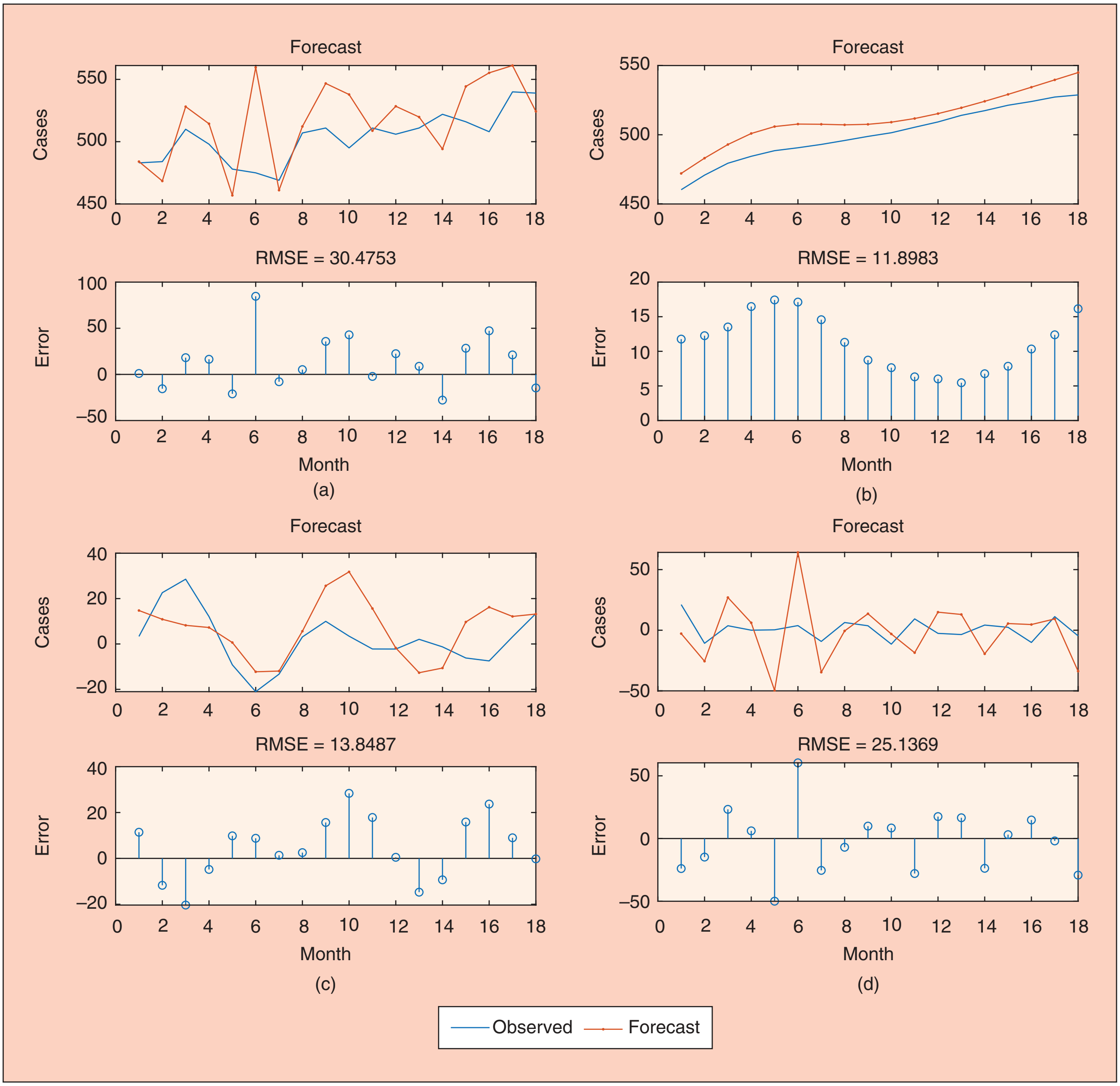


FIGURE 18 Predicted value and Root mean square error (RMSE) of the number of COVID-19 cases forecast corresponding to each IMF of Gujarat CC signal conducted on the test set (MAPE = 4.68%). (a) CC; (b) CC-IMF$_1$; (c) CC-IMF$_2$; and (d) CC-IMF$_3$.

in three states of India is proposed. Based on the literature review, the number of confirmed cases of COVID-19 is correlated with both temperature and humidity [8]–[11]. Therefore, both of these meteorological parameters are considered as features in training RNNs. Also, an equation proposed in [20] is used to calculate the transmission rates corresponding to each lockdown phase. As such, temperature, humidity and transmission rate have been used as features in this paper.

We conclude that specifying a soft transmission rate by smoothing the obtained step function can improve the prediction results. Moreover, compatible modes of signals were systematically derived, and it was found that training those with similar center frequency in separate RNNs improved the predictions. We collected the information from both outbreaks and available meteorological parameters to construct a model for predicting the future number of confirmed cases of COVID-19. However, one needs to take the following into account when predicting the future occurrence of COVID-19 using the proposed model:

1) The future value for transmission rates corresponding to a set of plausible lockdown phases may be approximated as those obtained from the previous lockdown stages.
2) The forecast value of temperature and humidity are usually available for some successive following days and can be used as features in the trained RNNs.

We have also shown through decomposing CC, T, and H signals into their modes using VMD that there are similar modes with close center frequencies in all of these signals. Although this confirms the effect of the temperature and humidity on the number of confirmed cases, one needs to look more carefully into the phase of the similar modes to unfold these dependencies more systematically. This issue contains sufficient merit to warrant independent research and can be a subject of future work.

Finally, the proposed procedure can provide insight into systematically forecasting the future number of COVID-19 cases, considering other factors affecting its spread in the community, which may include health policy, mask usage rate, and wind speed. As the method has been shown to be successful when applied to different Indian states that have quite different meteorological dynamics, it could be applied to other countries, especially if extended to additional relevant factors. The method proposed in this paper can also be used for other time series forecasting problems when complex signals are used as features.

References

[1] C. Huang et al., "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China," *Lancet*, vol. 395, no. 10223, pp. 497–506, 2020. doi: 10.1016/S0140-6736(20)30183-5.

[2] L. Wang, Y. Wang, D. Ye, and Q. Liu, "A review of the 2019 novel coronavirus (COVID-19) based on current evidence," *Int. J. Antimicrob. Agent*, p. 105,948, 2020. doi: 10.1016/j.ijantimicag.2020.105948.

[3] C. I. Jarvis et al., "Quantifying the impact of physical distance measures on the transmission of COVID-19 in the UK," *BMC Med.*, vol. 18, p. 124, 2020. doi: 10.1186/s12916-020-01597-8.

[4] H. Lau et al., "The positive impact of lockdown in Wuhan on containing the COVID-19 outbreak in China," *J. Travel Med.*, vol. 27, no. 3, p. taaa037, 2020. doi: 10.1093/jtm/taaa037.

[5] F. E. Alvarez, D. Argente, and F. Lippi, "A simple planning problem for covid-19 lockdown," National Bureau of Economic Research, Tech. Rep., 2020.

[6] S. Gupta, G. S. Raghuwanshi, and A. Chanda, "Effect of weather on COVID-19 spread in the US: A prediction model for India in 2020," *Sci. Total Environ.*, p. 138,860, 2020. doi: 10.1016/j.scitotenv.2020.138860.

[7] H. Qi et al., "COVID-19 transmission in Mainland China is associated with temperature and humidity: A time-series analysis," *Sci. Total Environ.*, p. 138,778, 2020. doi: 10.1016/j.scitotenv.2020.138778.

[8] B. Oliveiros, L. Caramelo, N. C. Ferreira, and F. Caramelo, "Role of temperature and humidity in the modulation of the doubling time of COVID-19 cases," *medRxiv*, 2020.

[9] J. Wang, K. Tang, K. Feng, and W. Lv, "High temperature and high humidity reduce the transmission of COVID-19," 2020.

[10] A. Auler, F. Cássaro, V. da Silva, and L. Pires, "Evidence that high temperatures and intermediate relative humidity might favor the spread of COVID-19 in tropical climate: A case study for the most affected Brazilian cities," *Sci. Total Environ.*, p. 139,090, 2020. doi: 10.1016/j.scitotenv.2020.139090.

[11] J. Demongeot, Y. Flet-Berliac, and H. Seligmann, "Temperature decreases spread parameters of the new COVID-19 case dynamics," *Biology*, vol. 9, no. 5, p. 94, 2020. doi: 10.3390/biology9050094.

[12] Y. Ma et al., "Effects of temperature variation and humidity on the death of COVID-19 in Wuhan, China," *Sci. Total Environ.*, p. 138,226, 2020. doi: 10.1016/j.scitotenv.2020.138226.

[13] A. Tomar and N. Gupta, "Prediction for the spread of COVID-19 in India and effectiveness of preventive measures," *Sci. Total Environ.*, p. 138,762, 2020. doi: 10.1016/j.scitotenv.2020.138762.

[14] D. Fanelli and F. Piazza, "Analysis and forecast of COVID-19 spreading in China, Italy and France," *Chaos, Solitons Fractal*, vol. 134, p. 109,761, 2020. doi: 10.1016/j.chaos.2020.109761.

[15] S. L. Chang, N. Harding, C. Zachreson, O. M. Cliff, and M. Prokopenko, "Modelling transmission and control of the COVID-19 pandemic in Australia," 2020, arXiv:2003.10218.

[16] R. Salgotra, G. Mostafa, and A. H. Gandomi, "Time series analysis and forecast of the COVID-19 pandemic in India using genetic programming," *Chaos, Solitons Fractal*, p. 109,945, 2020. doi: 10.1016/j.chaos.2020.109945.

[17] R. Salgotra and A. H. Gandomi, "Time series analysis of the COVID-19 pandemic in Australia using genetic programming," in *Data Science for COVID-19*. Amsterdam, The Netherlands: Elsevier, 2020.

[18] T. Sardar, S. S. Nadim, and J. Chattopadhyay, "Assessment of 21 days lockdown effect in some states and overall India: A predictive mathematical study on COVID-19 outbreak," 2020, arXiv:2004.03487.

[19] R. Salgotra, S. Singh, U. Singh, S. Saha, and A. H. Gandomi, "COVID-19: Time series datasets India versus World," 2020.

[20] C. Kirkeby, T. Halasa, M. Gussmann, N. Toft, and K. Græsbøll, "Methods for estimating disease transmission rates: Evaluating the precision of Poisson regression and two novel methods," *Sci. Rep.*, vol. 7, no. 1, pp. 1–11, 2017. doi: 10.1038/s41598-017-09209-x.

[21] D. Garcia, "Robust smoothing of gridded data in one and higher dimensions with missing values," *Comput. Statist. Data Anal.*, vol. 54, no. 4, pp. 1167–1178, 2010. doi: 10.1016/j.csda.2009.09.020.

[22] K. Zolna, P. B. Dao, W. J. Staszewski, and T. Barszcz, "Towards homoscedastic nonlinear cointegration for structural health monitoring," *Mech. Syst. Signal Process.*, vol. 75, pp. 94–108, 2016. doi: 10.1016/j.ymssp.2015.12.014.

[23] P. B. Dao and W. J. Staszewski, "Data normalisation for Lamb wave–based damage detection using cointegration: A case study with single-and multiple-temperature trends," *J. Intell. Mater. Syst. Struct.*, vol. 25, no. 7, pp. 845–857, 2014. doi: 10.1177/1045389X13512186.

[24] D. Kwiatkowski et al., "Testing the null hypothesis of stationarity against the alternative of a unit root," *J. Econom.*, vol. 54, nos. 1–3, pp. 159–178, 1992. doi: 10.1016/0304-4076(92)90104-Y.

[25] W. K. Newey and K. D. West, "A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix," National Bureau of Economic Research, Tech. Rep., 1986.

[26] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 531–544, 2014. doi: 10.1109/TSP.2013.2288675.

[27] D. Zosso, "Variational mode decomposition," Matlab Central File Exchange. Accessed: June 22, 2020. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/44765-variational-mode-decomposition

[28] Z. Zhao, W. Chen, X. Wu, P. C. Chen, and J. Liu, "LSTM network: A deep learning approach for short-term traffic forecast," *IET Intell. Transport Syst.*, vol. 11, no. 2, pp. 68–75, 2017. doi: 10.1049/iet-its.2016.0208.

[29] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-term residential load forecasting based on LSTM recurrent neural network," *IEEE Trans. Smart Grid*, vol. 10, no. 1, pp. 841–851, 2017. doi: 10.1109/TSG.2017.2753802.

[30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

[31] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, arXiv:1412.6980.

[33] D. Gabor, "Theory of communication. Part 1: The analysis of information," *J. Inst. Elect. Eng. III, Radio Commun. Eng.*, vol. 93, no. 26, pp. 429–441, 1946. doi: 10.1049/ji-3-2.1946.0074.