# Prediction of Sea Surface Temperature Using Long Short-Term Memory

Qin Zhang, Hui Wang, Junyu Dong, Member, IEEE, Guoqiang Zhong, Member, IEEE, and Xin Sun, Member, IEEE

伊俊杰

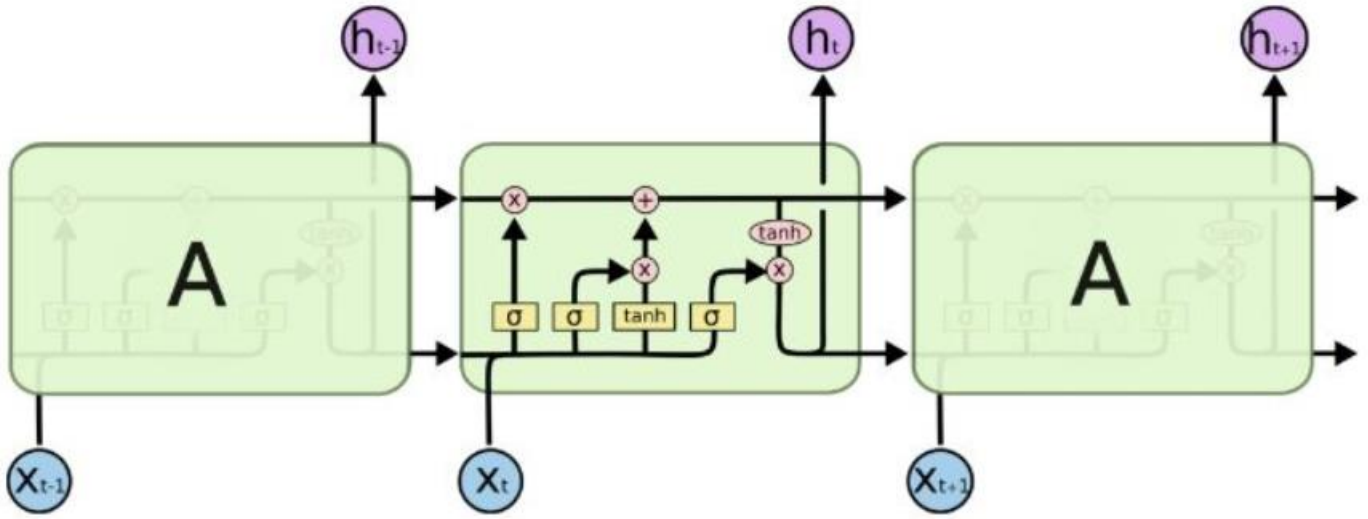
# Background

- The former tries to utilize a series of differential equations to describe the variation of SST, which is usually sophisticated and demands increasing computational effort and time. In addition, numerical model differs in different sea areas
- SST的变化 通常是复杂的,需要增加计算量和时间, 不同海域的数值模型也不同,因此很难采用传统物理学的数值模型来预测
- Long short-term memory is a special kind of recurrent neural network (RNN), which is a class of artificial neural network where connections between units form a directed cycle. However, vanilla RNN suffers a lot about vanishing or exploding gradient prob- lem, which cannot solve the long-term dependence problem. And it is very difficult to train. While LSTM introduces the gate mechanism to prevent back-propagated errors from vanishing or exploding, which has been subsequently proved to be more effective than conventional RNNs
- 长短期记忆是一种特殊的循环神经网络(RNN),它是一类人工神经网络,其中单元之间的连接形成有向循环,但是RNN在梯度消失或梯度爆炸遭受了很多困扰,无法解决长期依赖问题,而且训练非常困难。LSTM引入了门机制来防止反向传播的梯度消失或梯度爆炸, 比传统RNN更有效

# Problem Formulation

- Usually, the sea surface can be divided into grids according to the latitude and longitude. Each grid will have a value at an interval of time. Then the SST values can be organized as 3-D grids.
- 通常，海面可以根据纬度和经度分为网格。每个网格在时间间隔上都有一个值。然后，可以将SST值组织为三维网格。
- Suppose the SST values from one single grid is taken during the time, it is a sequence of real values. If a model can be built to capture the temporal relationship among data, then the future values can be predicted according to the historical values. Therefore, the prediction problem at this single grid can be formulated as a regression problem: if k days' SST values are given, what are the SST values for the k+1 to k+l days? Here, l represents the length of prediction.
- 假设在此期间从一个网格中获取了 SST值，它是一连串的实际值。如果可以建立一个模型来捕获数据之间的时间关系，则可以根据历史值来预测将来的值。因此，可以将单个网格上的预测问题表述为回归问题：如果给出了k天的SST值，那么我们可以预测k + 1tok + l天的SST值

# LSTM

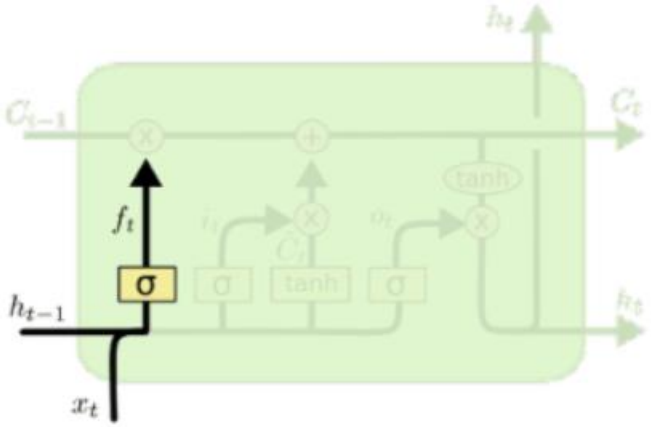



The repeating module in an LSTM contains four interacting layers.

RNN与LSTM最大的区别在于LSTM中最顶层多了一条名为“cell state”的信息传送带，其实也就是长期记忆，包含了之前时刻所有的有用的信息.所以cell state也可以理解为传送带，是整个模型中的记忆空间，随着时间而变化的.

LSTM中有3个控制门：输入门，输出门，记忆门,每一个门都有自己的参数矩阵,控制门的结构主要由一个sigmoid函数跟点乘操作组成；sigmoid函数的值为0-1之间，点乘操作决定多少信息可以传送过去，当为0时，不传送，当为1时，全部传送.

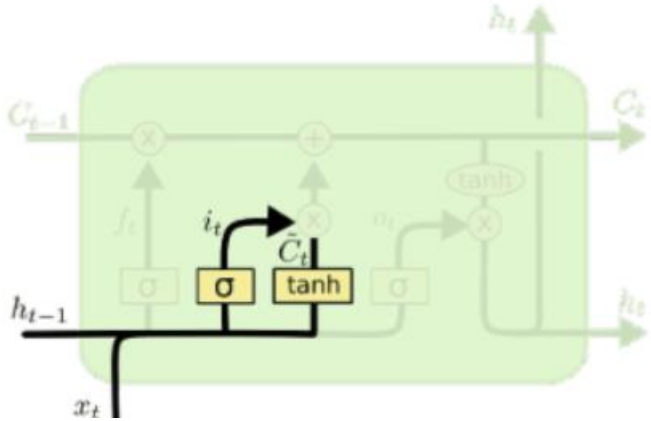
# LSTM-遗忘门

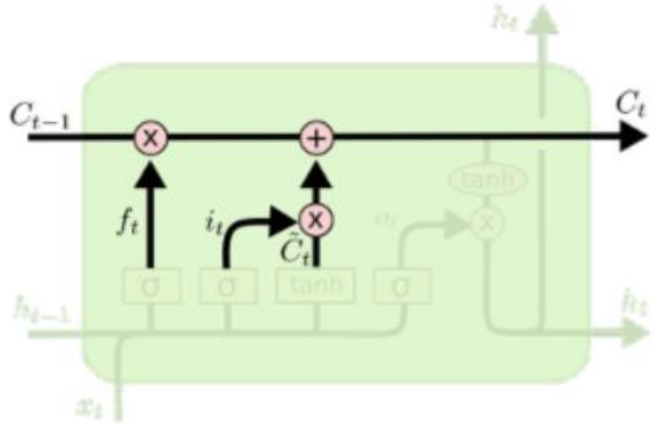



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

在我们 LSTM 中的第一步是决定我们会从cell state中丢弃什么信息。这个决定通过一个称为 **遗忘门** 完成。遗忘门会读取$h_{t-1}$和$x_t$，然后通过非线性激活函数得到一个0-1的值,将这个0-1的值跟传过来的$C_{t-1}$相乘,控制$C_{t-1}$输出多少,1 表示“完全保留”,0 表示“完全舍弃”

# LSTM-输入门



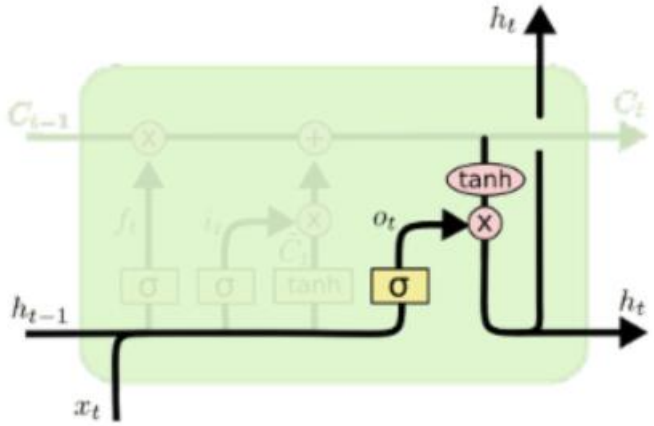
$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

**输入门**是决定哪些新的信息将被存储到cell state状态中，这包含两个部分。首先，一个叫“输入门层”的sigmoid层决定我们将要更新哪些值；然后，一个tanh层创建一个新的可被加入cell state的候选值向量$C_t$~.将上面得到的0-1的值和$C_t$~相乘,可以控制需要输入的大小,更新状态$C_t$。 1 表示"完全保留",0 表示"完全舍弃"
最后我们将旧状态$C_{t-1}$乘上$f_t$,先忘记我们决定忘记的信息。然后加上$i_t$*$C_t$~，即用每个状态值更新率$i_t$乘上候选值向量。

# LSTM-输出门

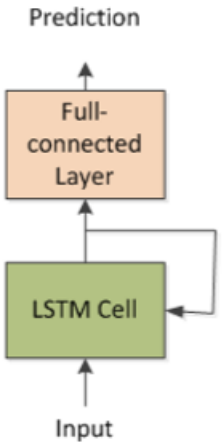



**输出门**需要决定输出是什么，这个输出将cell state，但需要被过滤一下。首先，我们用一个sigmoid门决定cell state的哪一部分将被输出；然后，将cell state经过一个tanh层（使得值被规范化到-1到1之间）；最后，将得到的值乘上sigmoid门输出的结果，从而得到我们决定输出的部分。

$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

# Basic LSTM Blocks

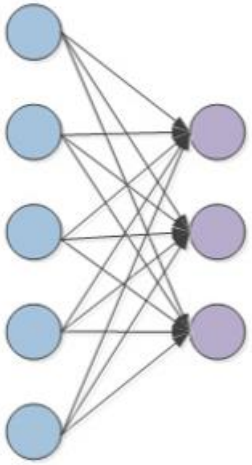



- LSTM is combined with a full-connected layer to build a basic LSTM block. Fig. 2shows the structure of a basic LSTM block. There are two basic neural layers in a block. The LSTM layer can capture the temporal relationship,i.e., the regular variation among the time series SST values. While the output of the LSTM layer is a vector i.e., the hidden vector of the last time step, a full-connected layer is used to make a better abstraction and combination for the output vector, and reduces its dimensionality,meanwhile maps the reducedvector to the final prediction. Fig. 3 shows a full-connectedlayer. The computation can be defined as follows:

$$(h', m') = \mathrm{LSTM}\left(\begin{bmatrix} Ix_i \\ h_{i-1} \end{bmatrix}, m, W\right)$$

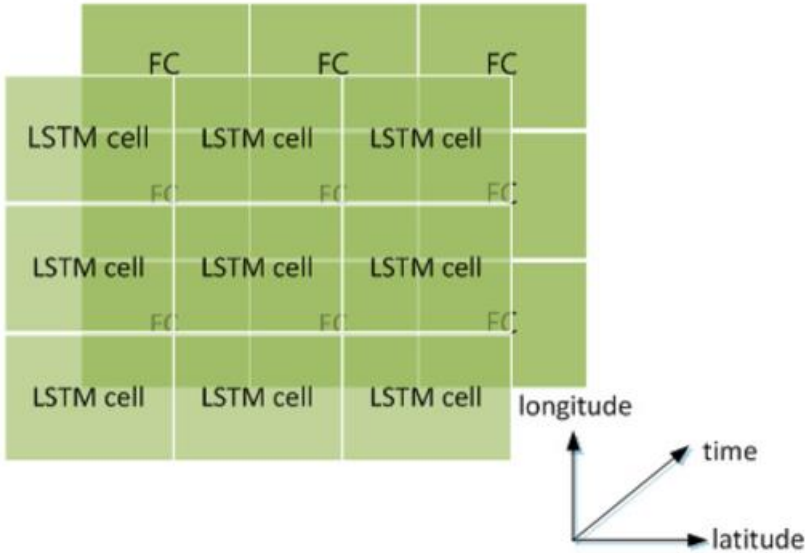
$$(h_i, m_i) = \mathrm{LSTM}\left(\begin{bmatrix} \mathrm{Iinput} \\ h_{i-1} \end{bmatrix}, m, W\right)$$

$$\mathrm{prediction} = \sigma(W^{\mathrm{fc}} h_l + b^{\mathrm{fc}})$$



- LSTM 与全连接层结合在一起以构建基本的 LSTM 模块。图 2 显示了基本 LSTM 块的结构。一个块中有两个基本的神经层。LSTM 层可以捕获时间关系，即时间序列 SST 值之间的规则变化。LSTM 层的输出是向量，即最后一步的隐藏向量，而全连接层用于对输出向量进行更好的抽象和组合，并降低其维数，同时映射缩小的向量最终预测。图 3 示出了全连接层。

# Network Architecture




This kind of block can predict future SST of a single grid, according to all the previous SST values of this grid. But it is still not enough. Prediction of SST of an area is needed. So the basic LSTM blocks can be assembled to construct the whole network.

根据该网格的所有先前SST值，此类块可以预测单个网格 的未来SST。但这还不够。需要预测一个区域的SST。因 此，可以组装基本的LSTM模块以构建整个网络

Fig. 4 shows the architecture of the network. It is like a cuboid: the x-axis stands for latitude, the y-axis stands for longitude, and the z-axis is time direction. Each grid corresponds to a grid in real data. Actually, the grids in the same place along the time axis form a basic block. We omit the connections between layers for clarity

图显示了网络的体系结构。就像一个长方体：x轴代表纬度，y轴代表经度，z轴代表时间方向。每个网格对应于实际数据中的网格。实际上，沿时间轴在相同位置的网格形成一个基本块。为了清楚起见，我们省略了层之间的连接 。

# Experimental Setup

- Since the SST prediction is formulated as a sequence prediction problem, i.e., using previous observations to predict the future, the duration the previous observations are to be used to predict the future should be determined. Of course, the longer the length is, the better the prediction will be. Meanwhile,morecomputationwill be needed.Here, the length of the previous sequence is set to four times of the length of prediction according to the characteristics of the periodical change of temperature data. In addition, there are still other important values to be determined: the number of layers for the LSTM layer lr and the full-connected layer lfc, which will determine the whole structure of the network. Also the corresponding number of hidden units denoted by units_r should be determined together.

- Once the structure of the network is determined, there are still other critical things to be determined in order to train the network, i.e., the activation function, the optimization method, the learning rate, the batch size, and so on. The basic LSTM block uses logistic sigmoid and hyperbolic tangent as an activation function. Here, we use an ReLU activation function for it is easy to optimize and is not saturated. The traditional optimization method for a deep network is stochastic gradient descent (SGD), which is the batch version of gradient descent. The batch method can speed up the convergence of network training. Here, we adopt the Adagrad optimization method [11], which can adapt the learning rate to the parameters, performing larger updates for infrequent and smaller updates for frequent parameters. Dean et al. [12] have found that Adagrad improved the robustness of SGD greatly and used it for training large-scale neural networks. We set the initial learning rate as 0.1, and the batch size as 100 in the following experiments.

# RESULTS AND DISCUSSION

TABLE I

PREDICTION RESULTS (RMSE) ON FIVE LOCATIONS WITH DIFFERENT *Units_rs*

| *units_r* | $p_1$ | $p_2$ | $p_3$ | $p_4$ | $p_5$ |
|---|---|---|---|---|---|
| 1 | 0.1595 | 0.1171 | 0.2690 | 0.2988 | 0.2626 |
| 2 | 0.1589 | 0.1137 | 0.2569 | 0.2909 | 0.2695 |
| 3 | 0.2075 | 0.0923 | 0.2580 | 0.2819 | 0.2606 |
| 4 | 0.2152 | 0.0918 | 0.2349 | 0.2752 | 0.2672 |
| 5 | **0.1280** | **0.0914** | **0.2310** | 0.2723 | **0.2362** |
| 6 | 0.1353 | 0.0922 | 0.2454 | **0.2646** | 0.2468 |

TABLE II

PREDICTION RESULTS (RMSE) ON FIVE LOCATIONS WITH DIFFERENT $l_r$s

| $l_r$ | $p_1$ | $p_2$ | $p_3$ | $p_4$ | $p_5$ |
|---|---|---|---|---|---|
| 1 | **0.1280** | **0.0914** | **0.2310** | **0.2723** | **0.2362** |
| 2 | 0.1288 | 0.1153 | 0.2500 | 0.2730 | 0.2496 |
| 3 | 0.3659 | 0.0950 | 0.2656 | 0.2732 | 0.3334 |

TABLE III

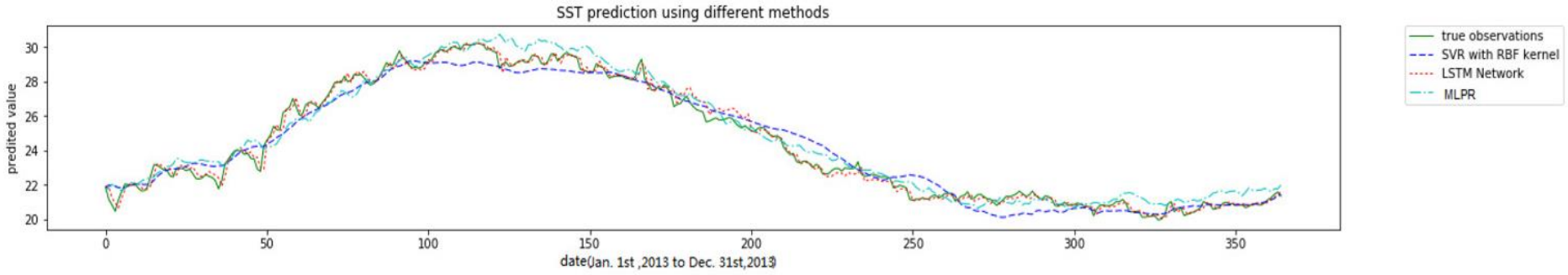
PREDICTION RESULTS (RMSE) ON FIVE LOCATIONS WITH DIFFERENT *k*s

| $l_{fc}$ | $p_1$ | $p_2$ | $p_3$ | $p_4$ | $p_5$ |
|---|---|---|---|---|---|
| 1[3] | **0.1280** | **0.0914** | **0.2310** | **0.2723** | **0.2362** |
| 2[3,3] | 0.2838 | 0.0945 | 0.2880 | 0.2724 | 0.2461 |
| 2[6,7] | 0.3660 | 0.2605 | 0.4655 | 0.2730 | 0.3355 |

TABLE IV

PREDICTION RESULTS (AREA AVERAGE RMSE) ON THE BOHAI SST DATA SET

| Methods | Daily | | Weekly | Monthly |
|---|---|---|---|---|
| | one day | three days | one week | one month |
| SVR | 0.3998 | 0.6158 | 0.4716 | 0.6538 |
| MLPR | 0.6633 | 0.8215 | 0.6919 | 0.8360 |
| LSTM network | **0.0767** | **0.1775** | **0.3844** | **0.3928** |



# 谢谢聆听

# THANK YOU