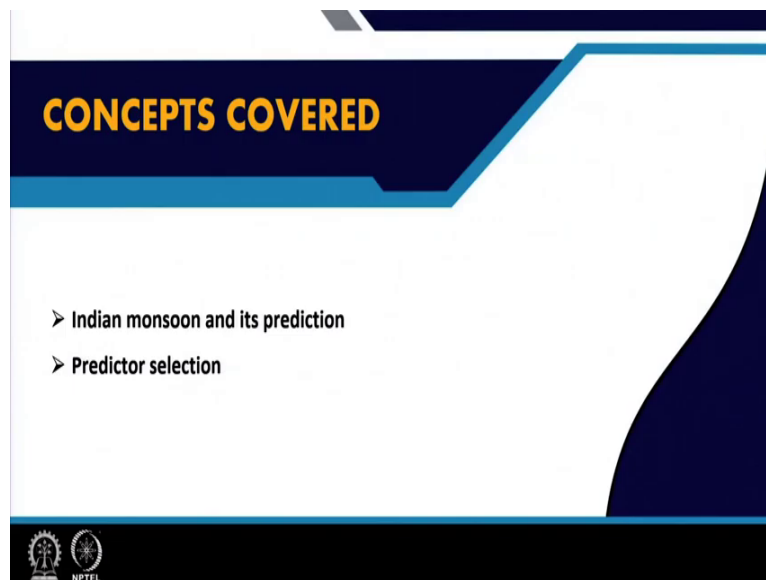


**Machine Learning for Earth System Sciences**  
**Prof. Adway Mitra**  
**Department of Computer Science and Engineering**  
**Centre of Excellence in Artificial Intelligence**  
**Indian Institute of Technology, Kharagpur**

**Module - 03**  
**Machine Learning for Discovering New Insights**  
**Lecture - 16**  
**Identification of Indian Monsoon Predictors**

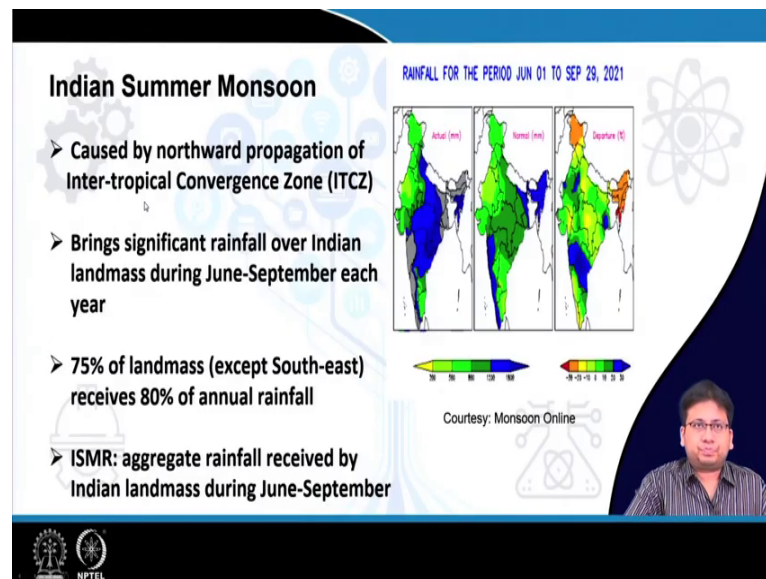
Hello everyone, welcome to lecture 16 of this course on Machine Learning for Earth System Science. Today we are beginning module 3. In this module we will be using Machine Learning to discover new insights about like the climate hydrology and various other domains of earth sciences.

(Refer Slide Time: 00:48)



The concepts to be covered in this lecture are Indian monsoon and its prediction and predictor selection.

(Refer Slide Time: 00:55)



So, like first of all let us talk about the background problem that is the Indian Summer Monsoon. So, we are all familiar with the Indian summer monsoon which is the usual monsoon season which covers most of India from June to September every year. So, this summer monsoon is caused by the northward propagation of the inter tropical convergence zone which is like a band of clouds which is usually near the equator, but during the summer it moves over the northern hemisphere and especially the tropical regions covering much of India.

Now this Indian summer monsoon which is actually part of the Asian summer monsoon which also includes lots of countries like in Eastern and sorry in Southern and South East Asian Asia this brings significant rainfall over the Indian land mass during June to September of each year except the South Eastern coast we know that parts of Andhra and Tamil, they receive the rainfall during the October and November period.

But most of the remaining landmass of India, it receives 75 % of the rain of their annual rain or I am sorry 80 % of their annual rainfall during these four months. The remaining 8 months that accounts for only about 20 % of their annual rainfall for most of the places. There are of course, other places like let us say the this western coast and this north eastern region which receives significant rainfall at other times also.

But large parts of India receives like nearly 80 % of the annual rainfall during the Indian summer monsoon. Now this ISMR, Indian Summer Monsoon Rainfall this is a quantity which refers to the aggregate rainfall received by the Indian land mass during the June to September period of any given year. So, ISMR like its defined for every year there have been various other definitions of ISMR.

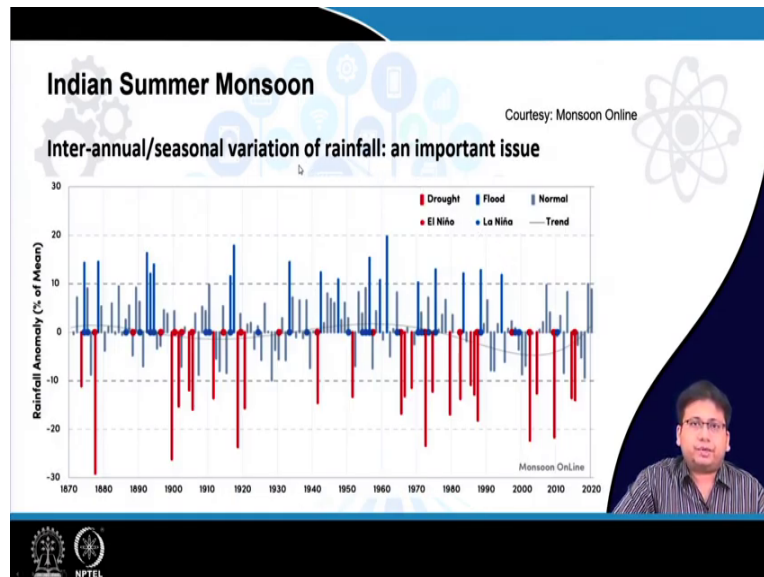
So, like earlier for example, like people used to may consider only what is known as the monsoon zone that is the covering the Central India as the as the to calculate the spatial aggregate of this monsoon rainfall, but nowadays most other studies they consider the entire Indian land mass when they are talking about the ISMR. Now there is a lot of spatial variation of this of in rainfall during monsoon; there are some regions which receive high rainfall and some regions which receive relatively less rainfall.

So, if you see the normal spatial distribution of rainfall by normal what I mean is the climatology the long period average. If you consider the 100 year rain let us say 100 year rainfall you will see that the these the western coast north east and these and some other regions here in the along the north eastern region, they receive a very high rainfall compared to the other parts.

But if you consider let us say the desert regions of Rajasthan or the these the southern coast near Tamil Nadu Andra etcetera they receive relatively less rainfall. So, this is of the long period average now if you consider like any particular year then they like each different regions can deviate differently. So, in a particular year it may happen that some regions receive more rainfall compared to the average while other regions receive less rainfall and so, on.

Now, but when we are talking about ISMR we are not talking about all these spatial variations we are just considering the spatial aggregate.

(Refer Slide Time: 04:57)

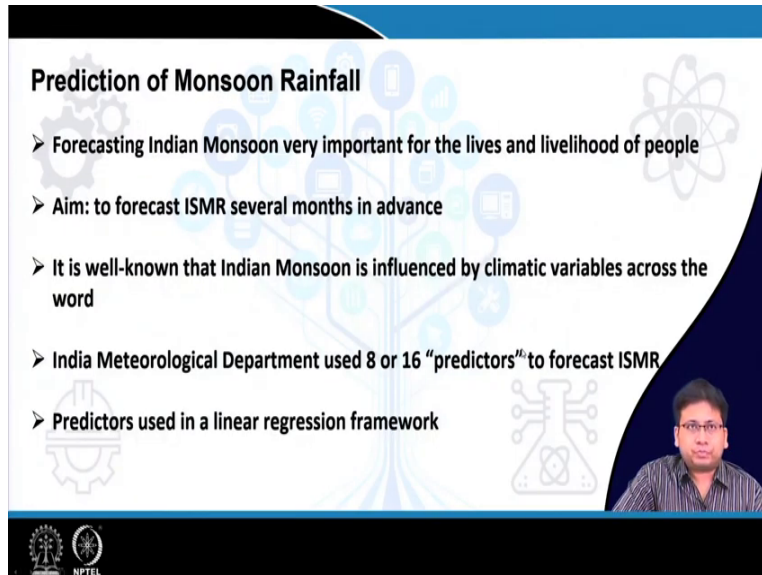


Now, the like the other variation which is associated with the like with the Indian monsoon is in the temporal domain. Now such temporal variation that is either intra seasonal that is within a season and as well as inter seasonal or inter annual that is across different seasons.

So, here this plot, it shows the anomaly of ISMR in different years. So, like the base case is of course, zero which means like just the long period average, but in every year it will the rain the overall aggregate ISMR recorded will be either slightly above or slightly below. Now, if the deviation is more than 10 % of the mean then such years are known as the excess rainfall years like you can see these blue towers like this.

And then those years where the rainfall is more than 10 % below the mean they are known as the deficient rainfall years. So, that is shown as this red lines which are jutting out below this dashed line. So, like on an average we most of the years are normal, but about 15 to 20 % of the years have excess rainfall and another 15 to 20 % years have deficient rainfall.

(Refer Slide Time: 06:20)



**Prediction of Monsoon Rainfall**

- Forecasting Indian Monsoon very important for the lives and livelihood of people
- Aim: to forecast ISMR several months in advance
- It is well-known that Indian Monsoon is influenced by climatic variables across the world
- India Meteorological Department used 8 or 16 “predictors” to forecast ISMR
- Predictors used in a linear regression framework

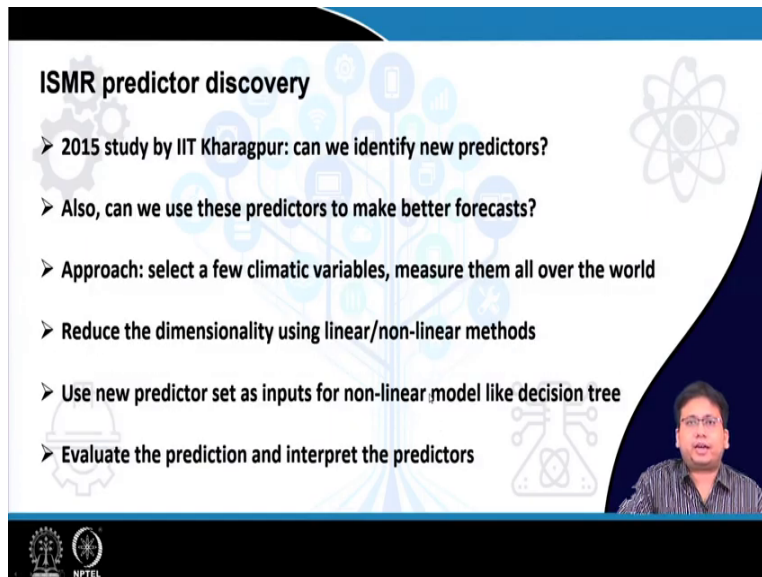
The slide features a blue and white color scheme with a background of faint icons related to weather and technology. A small video inset in the bottom right corner shows a man with glasses speaking. The NPTEL logo is visible in the bottom left corner.

Now, we all know that this that much of our agriculture and economic activities are dependent on the Indian monsoon. So, forecasting the Indian monsoon is very important for the lives and livelihood of the people. Often the India meteorological department tries to make forecast that this year there is going to be normal monsoon or this year there is going to be deficient monsoon etcetera so, that the various stakeholders they can come up with contingency measures that is if it is known that there is going to be less monsoon rainfall, then people will have the farmers will have to like change their cropping patterns accordingly they will have to plan alternative sources of irrigation since much of the irrigation comes from rainfall and so, on.

On the other hand, there if in any year we know that there is going to be excess rainfall then there is always a risk of floods, especially, in the low lying areas which we can call as flood prone. So, people have to take contingency measures and plan to evacuate if necessary. So, the it is very important to forecast the Indian monsoon ahead of time, especially several months in advance now how to do this forecast?

So, it is well known that Indian monsoon is influenced by climatic factors from all over all the over the world. So, earlier the India Meteorological Department or the IMD they used to have some 8 or 16 predictors from different parts of the world which they used as a in a simple linear regression framework to predict the Indian that the ISMR which we talked about.

(Refer Slide Time: 08:05)



**ISMR predictor discovery**

- 2015 study by IIT Kharagpur: can we identify new predictors?
- Also, can we use these predictors to make better forecasts?
- Approach: select a few climatic variables, measure them all over the world
- Reduce the dimensionality using linear/non-linear methods
- Use new predictor set as inputs for non-linear model like decision tree
- Evaluate the prediction and interpret the predictors

The slide features a background with faint icons of a gear, a lightbulb, a bar chart, and a network diagram. A small video inset in the bottom right corner shows a man with glasses speaking. The NPTEL logo is visible in the bottom left corner.

So, these 8 or 16 predictors they had been like identified over the years because the meteorological department has been active for more than a 100 years in India right from the colonial times and the climate scientists or meteorologists at that time based on their empirical understanding, they had identified various climatic variables from other parts of the world which might give us some indication of how the ISMR is going to be.

However, like those 8 or 16 predictors which we talked about those may have been like based on very empirical understanding of the meteorologist, but now like now that we have a vast amount of observations null data as well as we have huge computing power as well and the algorithms and models from machine learning can be somehow use this to make a more efficient forecast and in the process can we also identify more crucial predictors.

So, that is the aim of the modern the monsoon forecast in the recent times. So, in this lecture we will talk about a 2015 study which was carried out by some researchers from IIT Kharagpur where the aim was to predict monsoon the predict ISMR ahead of time and also identify the new predictors. So, the basic approach was like first of all select a few climatic variables and then measure these from all over the world at different locations and at different points of time.

And so, all of those things are going to be potential predictors now this set of potential predictors will have a lot of redundancy because many of them will themselves be dependent on each other. So, like as I said each climatic variables is going to be measured from all over the world because we do not know like which region is going to influence the Indian monsoon like we know that there is something known as tele connections based on by which even variables far away from India let us say from the pacific or the Atlantic Ocean, they also can have a bearing on Indian monsoon rainfall.

So, nowso, we have a very large pool of potential predictors, but many of them may themselves be dependent on each other ah. So, it is necessary to reduce their dimensionality using some linear or non-linear methods. Now linear dimensionality reduction is PCA, but we can also use non-linear methods for dimensionality reduction.

In the previous lecture we talked about auto encoders for example, now once we have done that once we have created a new set of predictors which are actually some like linear or non-linear combinations of the huge pool of predictors which we initially talked about next these new predictors are going to be used as inputs to some non-linear prediction model like a decision tree with whose output will be the desired quantity namely the ISMR of any given year ok.

And once we have done that we of course, need to evaluate the prediction and also interpret the predictors this last part is very important because in a domain like earth sciences it is not enough to somehow concoct some kind of a machine learning model which will be able to do the predictions correctly because we also want to understand the science behind the process, we also like we want that is why it is important to be able to interpret the results also.

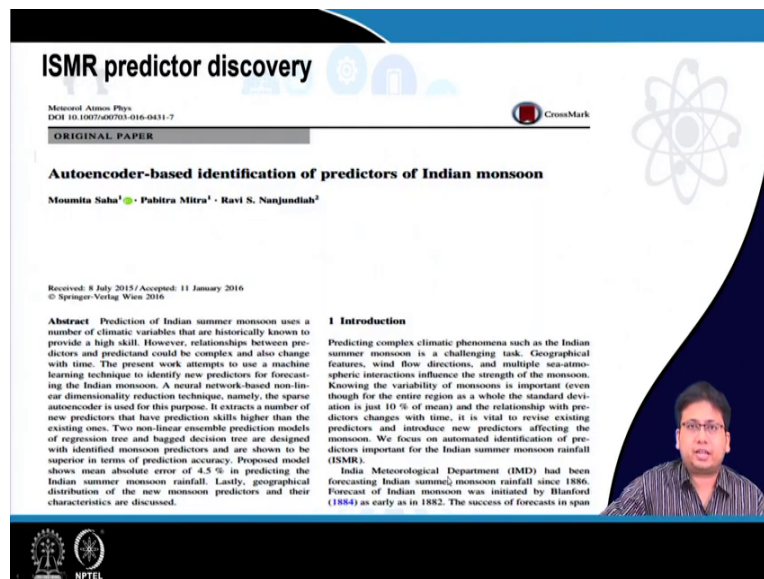
The climate scientists are like interested to gain new insights about the climatic process which we may they may be able to do if they are able to know what are the predictors of any given phenomena like Indian monsoon, but those predictors should also have some kind of sense; I mean like we cannot just say anything arbitrary like we have already that might happen because there is this concept of spurious correlations.

So, if you remember like in the lecture about causality we had talked about spurious correlations, we have seen that how two variables seemingly unconnected variables may be related due to the

presence of confounders and something like that. So, if we are not focusing on the on the necessary variables, we may end up identifying some kind of spurious predictors like that which will like which may be fine for forecasting purposes, but it will not help you to understand the process.

So, it is necessary that with the we be able to interpret the predictors that we are finding.

(Refer Slide Time: 13:05)



So, this is the paper which I just mentioned or its called auto encoder based identification of predictors of Indian monsoon, it appeared in the year 2016. So, the abstract of the paper is that prediction of Indian summer monsoon uses a number of climatic variables that are historically known to provide a high skill.

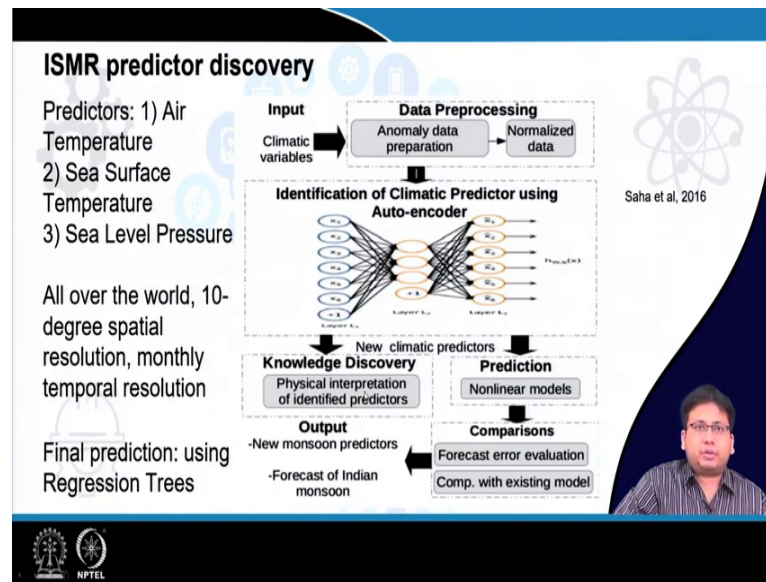
However, relationships between predictors and predictand could be complex and also change with time the present work attempts to use a machine learning technique to identify new predictors for forecasting Indian monsoon. A neural network based non-linear dimensionality reduction technique namely the sparse auto encoder is used for this purpose, it extracts a number of new predictors that have prediction skills higher than the existing ones.

Two non-linear ensemble prediction models of regression tree and bagged decision tree are designed with identified monsoon predictors and are shown to be superior in terms of prediction



accuracy. Proposed models shows mean absolute error of 4.5 % in predicting the Indian summer monsoon rainfall lastly geographical distributions of the new monsoon predictors and their characteristics are discussed.

(Refer Slide Time: 14:28)



So, this is the block diagram of this paper which we just discussed by Saha et al in 2016. So, the predictor variables which they have considered these are the air temperature, sea surface temperature and sea level pressure which have been measured from all over the world I mean of course, sea surface temperature and sea level pressure can be measured only over the seas not over the land.

So, they have but wherever these are defined they have measured it and they have considered 10-degree spatial resolution which is actually quite coarse grained (Refer Time: 15:04) compared to the standards at which we work today and we also they have also used monthly temporal resolution. So, the basic framework is like this first they have these input climatic variables which I just talked about.

Now, they do the data pre-processing by removing the anomaly value by calculating the anomalies removing the missing values and other may be noisy values and by normalizing the data, next as I said these input variables they may contain a lot of dependencies or as we say or

we can say redundancies. So, to remove that it is passed through like an auto encoder for the non-linear dimensionality reduction.

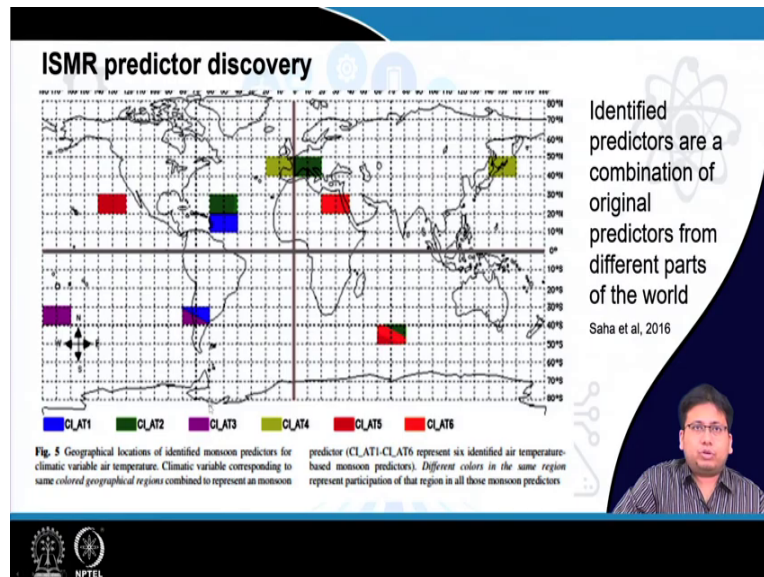
And what we get is basically the code which is a non-linear combination of these like of the original set of the predictors. Additionally, this auto encoder is a sparse auto encoder which means that most of these edge weights will be zero which means that the code which we are considering now each element of the code will be a non-linear function of only a small number of these input predictors of the original set of predictors.

This sparsity makes sure that like we do not end up using the predictors from all predictors from all over the world because that will make the process very complicated instead we learn to identify which are the most crucial ones and only those will be retained the less crucial ones they will just be ignored by setting their edge weights to 0.

So, once we have done that like ah. So, we once we have identified this new climatic predictors. So, first of all we can do the prediction using some kind of non-linear models like say in this case they have used decision trees and I mean ensemble of regression trees something like that and once they have the predictions, they can carry do the comparisons with the actual observations.

And then also quantify the forecast error and compare with the existing models which India Meteorological Department may have been using earlier the 8 parameter model 16 parameter models which we talked about. And then once also the these new predictors which we have identified as I said earlier it is very important to do a physical interpretation of this so, that we can get some idea about like the exact process in which that prediction is done.

(Refer Slide Time: 17:40)



So, the now if you look at the results. So, if we like as I said like the initial set of predictors were air temperature sea level sorry sea level pressure and sea surface temperature from all over the world now, especially if you consider at air temperature. So, as I said the what the auto encoder does is like it creates a code which is where each code each part of the code is like a non-linear combination of the input.

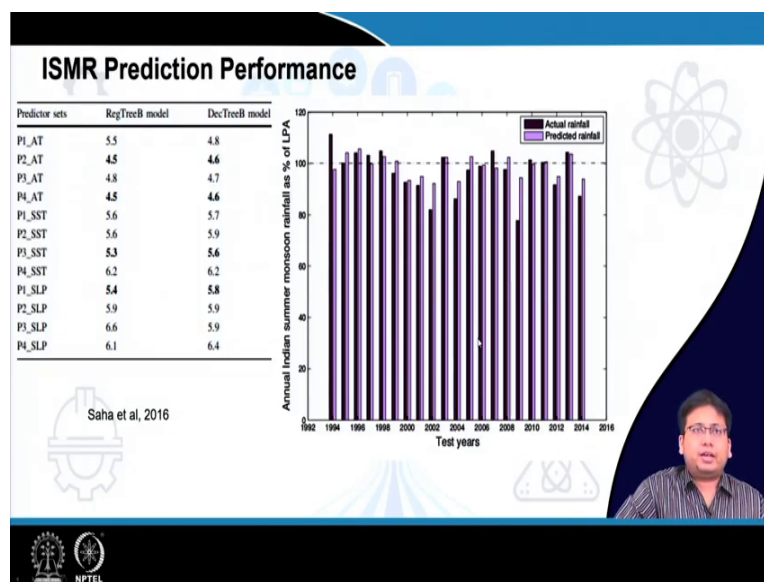
That is only sum of the input variables, they will be incorporated in a non-linear way to form one element of the code. So, like if we focus only on the air temperature part. So, there are six like we can say that the code has six new predictors each predictor being a non-linear combination of the original inputs. So, let us say what this figure shows is the geographical distribution of the different predictors that are used.

So, this red this indicates one code like one of the six codes. So, that the regions that are shown in red in this map these are the input original variables which are somehow combined non-linearly to come up with this this part of the code. Similarly, the like if you consider the third code like that is also made up from or made up of this the non-linear combination of the input predictors from or original input predictors from different parts of the world.

So, these violet colour you can see that these locations these are coloured in violet means that at the air temperature from all these locations these are non-linearly combined to form this new this predictor called this CI\_AT3. Then similarly for the blue predictor like there also its a non-linear combination of the air temperature from these regions as well as from these regions.

So, this shows the predicted or the identified predictors as a combination of I mean a non-linear combination of the original predictors from different part of the world this map focuses only on the air temperature, but then the other for the other variables sea surface temperature and sea level pressure for them also will get a map like this.

(Refer Slide Time: 20:19)



And now using all those things all those predictors we can actually carry out the prediction of the ISMR on of any given year using different models.

So, here they have considered the regression tree model as well as the ensemble of decision tree models and like what they are quantifying here is the error the that is the prediction error. So, like as you can understand the smaller the prediction error the better is the prediction. So, and here like we see the comparisons of like every year we see the comparisons of the actual ISMR against the what is predicted by this model.

There are a small number of years like for example, this 2009 where there is a large error of prediction, but then this 2009 was an exceptional year because there was a large drought caused by the aila cyclone, which had hit the eastern coast of India and that cyclone it appears that it had absorbed a lot of moisture due to which impacted the rain monsoon rainfall that year that is like that is an effect which is different from the from these long term effects or long range effects which we considered here.

# ISMR Prediction Performance

**Figure 11: Root Mean Square Error (RMSE) as % of LPA**

Model	RMSE (% of LPA)
IMD 16-param	~4.8
IMD 8-param	~3.8
IMD 10-param	~3.2
CI, AT	~2.8
CI, SST	~3.0
CI, SLP	~3.2

**Figure 12: Mean Absolute Error (MAE) as % of LPA**

Model	MAE (% of LPA)
IMD Opertional	~7.5
IMD LRF1	~7.2
IMD LRF2	~6.5
AT	~4.5
SST	~5.8
SLP	~4.5

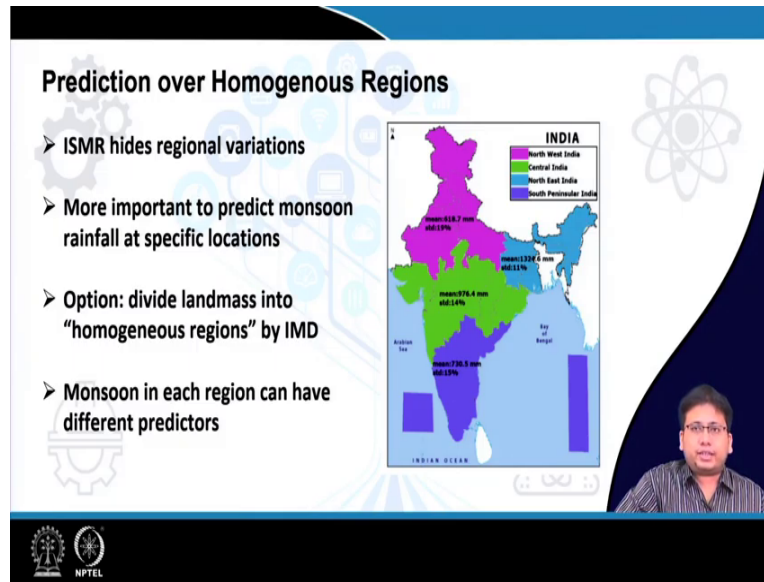
**Figure 13: Comparison of prediction by identified monsoon predictors of air temperature, CI, SST, and sea level pressure (CI, SST, and SLP) models with IMD existing operational (Rajeevan et al., 2004) and PPR (LRF1 and LRF2) (Rajeevan et al., 2000) models for period 2003–2014.**

Saha et al., 2016

Now, if you consider the performance of these predictors against the 8 parameter models or the 16 parameter models which have been used by IMD at different points of time or the various other operational models which IMD has used in recent times we see that these methods they give much less errors of forecast than the IMD methods. So, this can be an interesting improvement for the monsoon prediction.

The point to be noted here is that the predictors, which they have come up with here is quite different from the set of 8 or 16 predictors which I with the India meteorological department had earlier been using.

(Refer Slide Time: 22:55)



And now as I said the ISMR is like the spatial aggregate rainfall over the Indian landmass, but that might hide regional variations and it is more important to make predictions at specific locations rather than given aggregate value.

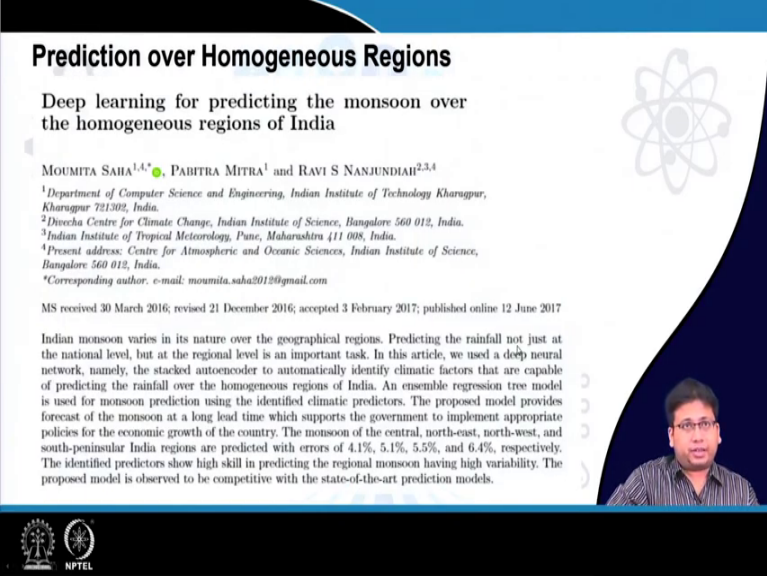
Because as I said or as I showed in the maps earlier the Indian region actually has significant heterogeneity regarding the monsoon rainfall. So, even if the overall ISMR on any given year is let us say is above average that does not mean that all the locations are going to have above average rainfall. In fact, some of the locations can have a below average or much below average rainfall also while some other regions may have very high rainfall and make up for that deficiency overall.

So, but these kinds of spatial variations are not taken into account when we are predicting the ISMR. So, a more meaningful way might be to predict the rainfall at more localized scales. So, by localized scales we may want to go down to grid level; now, predicting grid level rainfall is a

much bigger challenging problem, but simpler way might be to divide the Indian landmass into these kinds of homogeneous regions.

So, these homogeneous regions have been defined by the IMD, India Meteorological Department on the basis of the long period averages of rainfall. So, as you can see like North India is one region, the Central India is another region, the South India is another region the north east India is yet another region and so, on and so, forth. So, we can actually carry out the same process as earlier for each of these homogeneous regions separately.

(Refer Slide Time: 24:46)



## Prediction over Homogeneous Regions



Deep learning for predicting the monsoon over the homogeneous regions of India

MOUMITA SAHA<sup>1,4,\*</sup>, PAHTRA MITRA<sup>1</sup> and RAVI S NANJUNDIAH<sup>2,3,4</sup>

<sup>1</sup>Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, Kharagpur 721302, India.  
<sup>2</sup>Directorate Centre for Climate Change, Indian Institute of Science, Bangalore 560 012, India.  
<sup>3</sup>Indian Institute of Tropical Meteorology, Pune, Maharashtra 411 008, India.  
<sup>4</sup>Present address: Centre for Atmospheric and Oceanic Sciences, Indian Institute of Science, Bangalore 560 012, India.  
\*Corresponding author. e-mail: moumita.saha2012@gmail.com

MS received 30 March 2016; revised 21 December 2016; accepted 3 February 2017; published online 12 June 2017

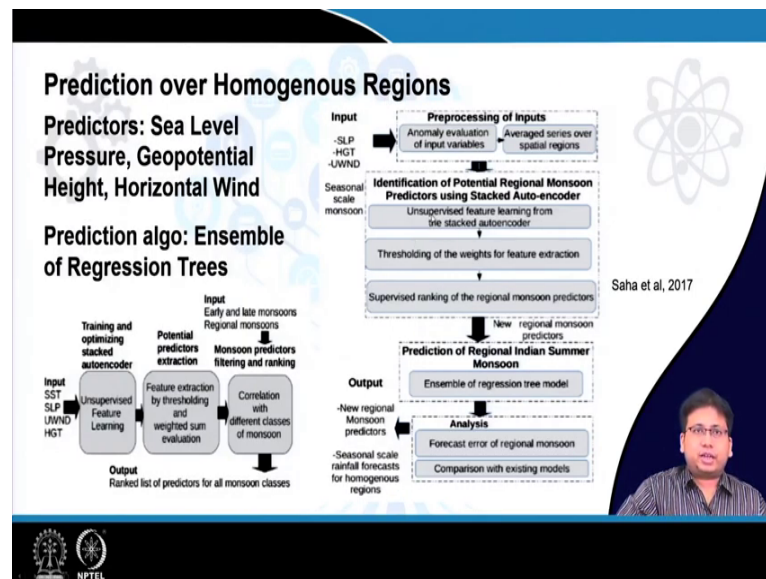
Indian monsoon varies in its nature over the geographical regions. Predicting the rainfall not just at the national level, but at the regional level is an important task. In this article, we used a deep neural network, namely, the stacked autoencoder to automatically identify climatic factors that are capable of predicting the rainfall over the homogeneous regions of India. An ensemble regression tree model is used for monsoon prediction using the identified climatic predictors. The proposed model provides forecast of the monsoon at a long lead time which supports the government to implement appropriate policies for the economic growth of the country. The monsoon of the central, north-east, north-west, and south-peninsular India regions are predicted with errors of 4.1%, 5.1%, 5.5%, and 6.4%, respectively. The identified predictors show high skill in predicting the regional monsoon having high variability. The proposed model is observed to be competitive with the state-of-the-art prediction models.



So, the same group of authors, they had a follow up paper to the previous paper where they actually do the prediction of over homogeneous regions using pretty much the same approach as earlier.



(Refer Slide Time: 25:00)



So, this time also they use the same set like same autoencoder approach for the non-linear dimensionality reduction and coming up with new predictor variables; however, in this case instead of air temperature sea surface temperature and sea level pressure they have a different set of input variables.

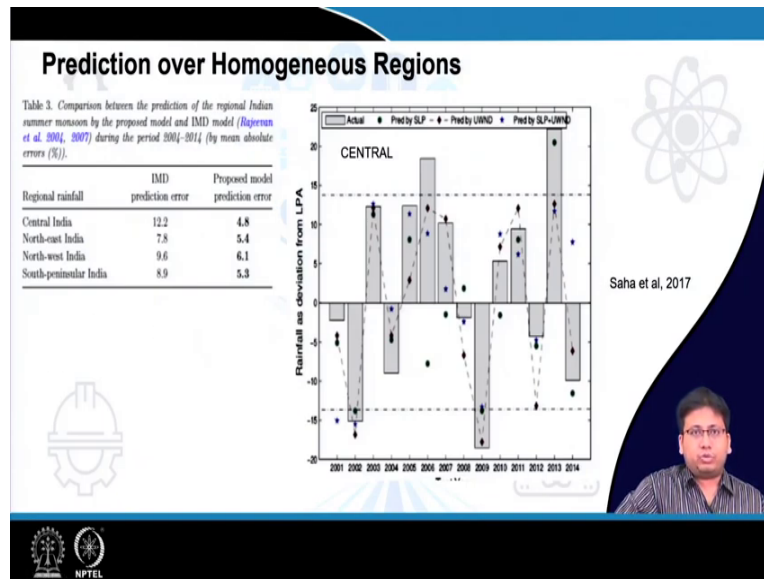
Namely sea level pressure of course, remains in this case, but they also consider the geopotential height and the UWND that is the horizontal wind like in different parts of the world and then like as in the previous case they like carry out they identified the anomalies of the input variables that is part of the pre-processing stage and once they have done so, they carry out the dimensionality reduction using the non-linear auto encoder which or in this case they use a slightly more advanced version of the auto encoder which is called the stacked auto encoder.

So, this again gives them the non-linear combination of these input features and then these input these new input the again the code of the auto encoder gives them like some new features which are combinations or sparse combinations of the original predictors and then those predictors are used using some kind of a classification or regression algorithm to predict the Indian summer monsoon rainfall at each of the different the homogeneous regions that we had talked about.



So, like the this process of this training and optimizing the auto encoder of calculating their edge weights etcetera like they like they use various strategies for that purpose which are all discussed in the paper. So, I request all of you or I strongly encourage all of you to read the paper and try to understand how they have done it.

(Refer Slide Time: 27:05)



So, these are some of the predictions they have obtained form for the. So, like in the unlike the previous case.


So, here for each of the different homogeneous regions they have to do the prediction separately and. So, they have reported the prediction errors also separately for each of the regions and as you can see the prediction error, they get by the proposed model is very significantly lesser than the what IMD has using their 8 parameter or 16 parameter models and others. So, like especially as you can see in case of central India there is a very drastic reduction of prediction error by this way .

So, similarly in the different years also like we you can see the comparison of the how good the like we like its also possible to do some kind of an ablation study like where instead of focusing on all the three variables they talked about the namely the UWND the geo potential height and

the sea level pressure they can instead of considering all of them together they can try like each of these things separately and see how strong the prediction would be in this case.

So, that kind of an ablation study they have done where they have actually compared each the predictions by each of those variables separately.

(Refer Slide Time: 28:34)

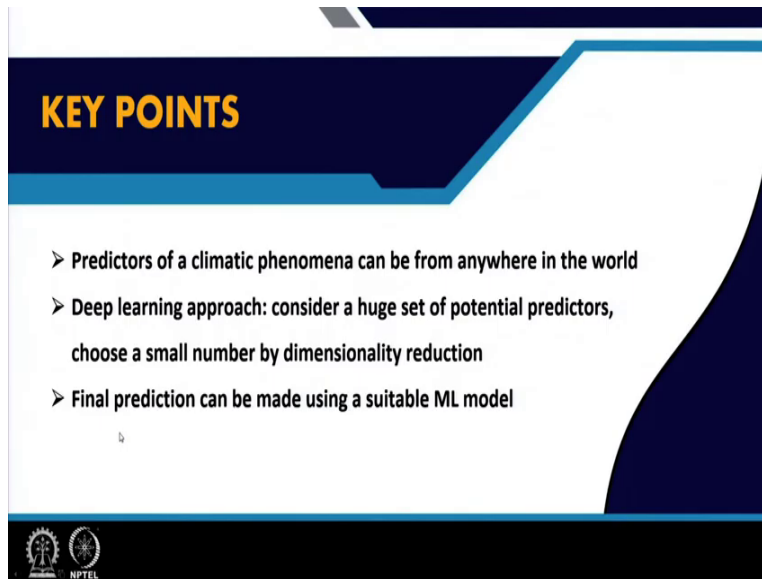


## REFERENCES

- Saha M, Mitra P, Nanjundiah RS. Autoencoder-based identification of predictors of Indian monsoon. *Meteorology and Atmospheric Physics*. 2016 Oct;128(5):813-28.
- Saha M, Mitra P, Nanjundiah RS. Deep learning for predicting the monsoon over the homogeneous regions of India. *Journal of earth system science*. 2017 Jun;126(4):1-8.

And so, basically these are the references which we considered for today. So, the first one was about ISMR prediction the second was about the rainfall annual rainfall prediction for each of the homogeneous regions both the papers are like basically on the using the same I approach which are summarized here.

(Refer Slide Time: 28:58)



**KEY POINTS**

- Predictors of a climatic phenomena can be from anywhere in the world
- Deep learning approach: consider a huge set of potential predictors, choose a small number by dimensionality reduction
- Final prediction can be made using a suitable ML model

NPTEL

First of all the predictors of the climatic phenomena can be from anywhere in the world. So, the deep learning approach is you just consider the huge set of potential predictors which might be the different climatic variables from all over the world and then we can choose a small number of them by some kind of a dimensionality reduction approach which can be done like using either linear methods like PCA or non-linear methods like auto encoder.

And once you have identified the new a new set of predictors, which are which might be some kind of combination of the original variables, then using those new predictors you may be able to do a with the prediction of the quantity of your interest be it ISMR or zone wise ISMR or something else also and that kind of final predictions you can make using a suitable machine learning model.

So, that brings us to the end of this lecture. So, this was our first case study on how machine learning methods can be used to gain new insights related to the earth sciences. So, the kind of insight we gained in this lecture is new predictors for Indian summer monsoon.

Of course, it is up to climate scientists to validate whether like how many whether all these predictors make sense to them as far as the physical process is concerned or not, but that is a different question.

So, the in fact, in these papers which we discussed the physical explanations of these predictors have also been discussed. So, the papers have been mentioned in the reference sections. I strongly urge you to go through both of these papers. So, in the next lecture we will consider a different use case or a different case study and. So, till then bye bye.