Prediction of temporal and spatial variability of daily precipitation using PCA and SOMs

Data project report for EOSC 510

By Mekdes Ayalew

December 15, 2017

Abstract

PCA and clustering methods are becoming commonly used methods to describe the main pattern of precipitation in different regions worldwide. The main purpose of this project is to replicate what is done before but using output from global climate models dynamically downscaled using regional climate model called WRF, in the Purcell mountain regions of British Columbia. For the downscaling approach, 80 km resolution ERA-Interim data is used as input to the WRF model, which is then downscaled to 10km resolution. PCA was performed both on the original and downscaled data. The PCA result from the downscaled data shows that the first two PCs explained 89% of the total variation and the first PC shows a pattern following topography. The first two significant WRF PCs modeled from the first five ERA-interim PCs shows strong significant correlation ($r = 0.78$) with the original PCs. The results from SOM shows different patterns where the most frequently occurring pattern (31%) involves low precipitation overall the domain which is distributed in a randomly manner. In conclusion, the results indicated that the main factor governing precipitation in our study area is topography and also modeled WRF PCs might be used for long-term projection of high-resolution temporal and spatial distribution of precipitation

1.Introduction

PCA and clustering methods are becoming commonly used methods to describe the main pattern of precipitation in different regions worldwide [Stathis (2009); Kiar Teo and others (2011); Othman and others (2015)]. These studies mainly used weather station data to determine the main pattern of precipitation in specific regions (Eg Spain, Malaysia). The main purpose of this project is to investigate different patterns of daily winter month's precipitation over the Purcell mountain regions of British Columbia over winter months of 2007 – 2017 using two methods such as Principal Component Analysis (PCA) and Self Organizing Map (SOM). Unlike the above studies, here we will be using reanalysis products from Global Climate Models dynamically downscaled using a regional climate model named Weather Research and Forecasting model (WRF). The motivation for applying dynamical downscaling is that regional climate models are generally thought to simulate the spatio temporal variability of precipitation more accurately than global climate models. The specific objectives of this project are :

1) To investigate the main pattern governing precipitation in winter months using the PCA approach.

2) To investigate the most characteristic features (temporal or spatial patterns) of winter daily precipitation using SOM.

3) To investigate the dominant patterns of precipitation over different range of values.

4) To derive a model that will be used to predict high-resolution temporal and spatial distribution of daily precipitation using results from PCA and SOM.

In the following sections we will briefly explain data and the preprocessing, results from PCA and SOM and finally discussion and conclusion.

2. DATA

2. 1 Data source and study site

The study site for this project is the Purcell mountain region of British Columbia where the center of the domain is Nordic glacier (Figure 1). Two sets of reanalysis data from ECMWF are used from which the first set of data is used as input for the WRF model and the second set of data is used as a comparison to the WRF model output. The full description of the two data sets is given in Table 1 and 2.

Figure 1: Purcell mountain regions of British Columbia

Table 1: ERA-Interim (ECMWF) data used as input to the WRF model

Source	ECMWF (ERA-Interim)
Temporal resolution	6 Hourly
Data type	surface and pressure level
Study period	Winter months (Dec, Jan and Feb) of 2007 -2017
Horizontal resolution	80 km
Number of vertical levels	60
Region /Area	North America

Table 2: Total daily precipitation from ERA-Interim (ECMWF)

Source	ECMWF (ERA-Interim)
Temporal resolution	3 Hourly
Data type	surface (total daily precipitation)
Study period	Winter months (Dec, Jan and Feb Feb) of 2007 - 2017
Horizontal resolution	80 km
Region	Purcell mountain region (Nordic glacier is the center of the domain)
Domain size	480 km x 240 km

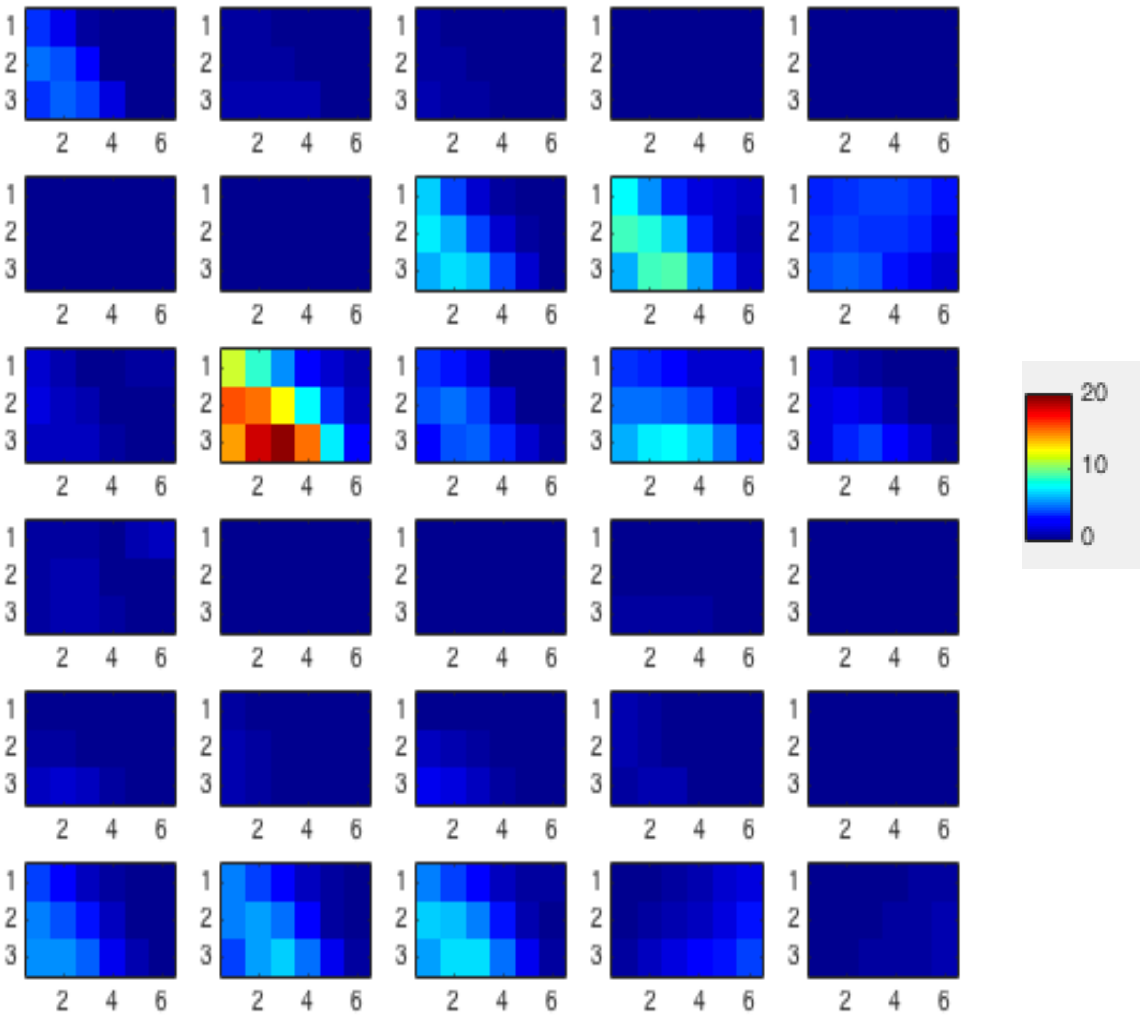


Figure 2: ERA-Interim, daily total precipitation on December 2010.

2.2 Data Preprocessing

The WRF model is used to dynamically downscale the ERA-Interim data. The model specification is given in Table 3. Figure 3 shows the downscaled daily total precipitation in December 2010, which provides more detailed patterns in comparison to ERA-interim data (Figure 2).

Table 3: Specification of the WRF model run

Model spin up	3 days
Temporal resolution	Daily
Data type	Daily total precipitation
Study period	Winter months (Dec, Jan and Feb) of 2007 - 2017 (10 years)
Horizontal resolution (grid size)	33 km (D01) and 10 km (D02)
Domain size	210 km × 210 km

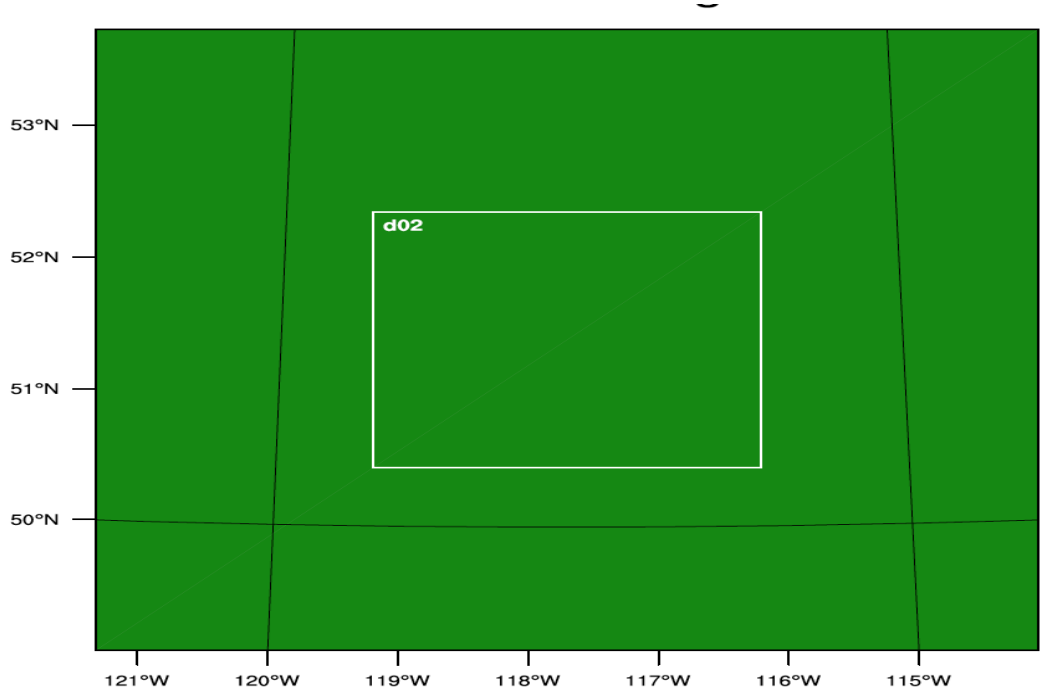


Figure 3: Domain set up (d01 – 33km) and (d02 – 10km).

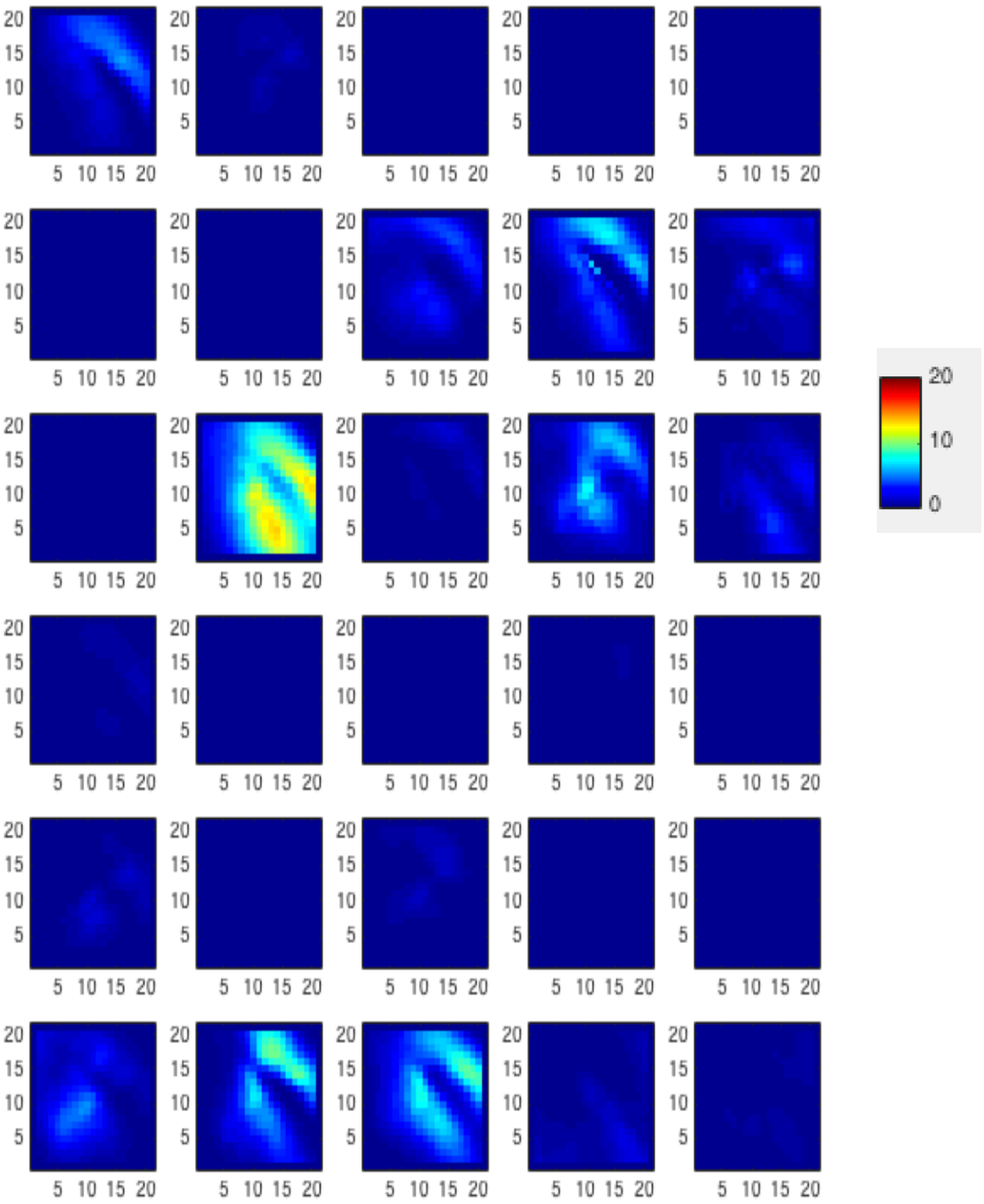


Figure 4: Daily total precipitation on December 2010 from WRF.

3. RESULTS

3.1 Results from PCA

We were not able to download ERA-interim data for exactly the same domain size as the WRF inner domain (10 km). Therefore, there is a slight difference in size and this might affect some of the results, especially when we do comparison.

3.1.1. PCA using the ERA-Interim data

PCA was applied on the ERA-interim data and the first three modes explain 99% of the total variance (Figure 5).

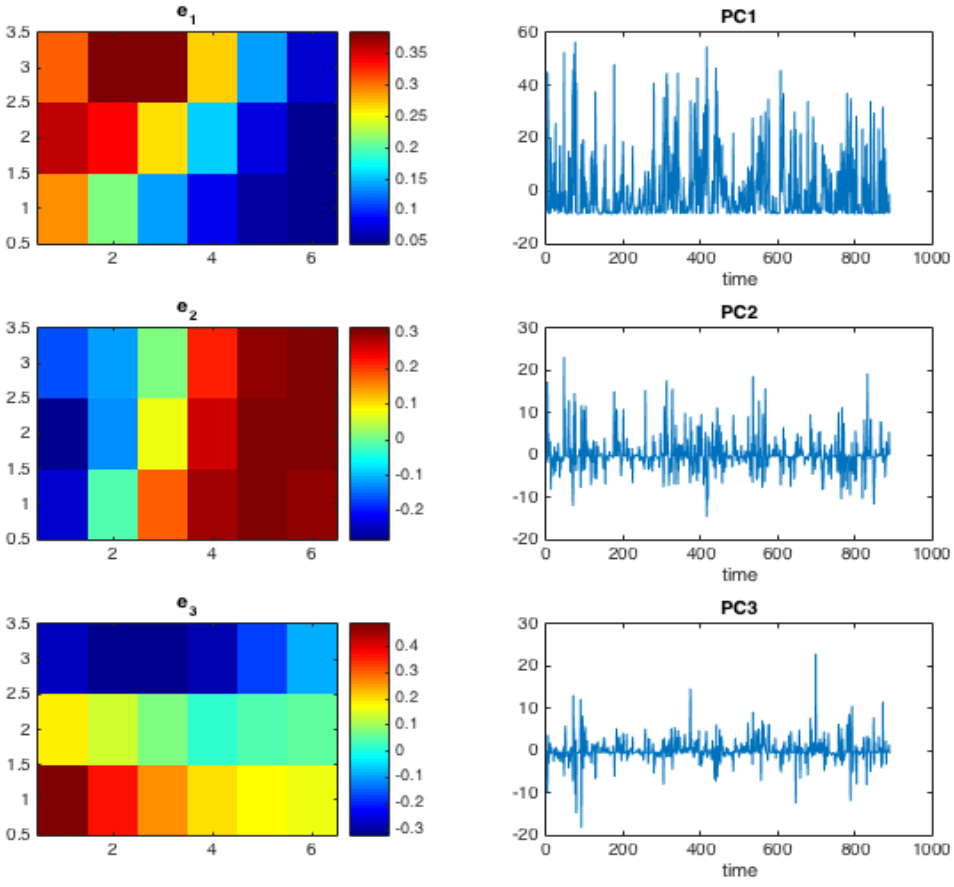


Figure 5: Significant eigen vectors and their principal components

3.1.2. PCA using the WRF model output

PCA was performed on the WRF output of daily precipitation data. The first four eigen vectors explain 94% of the total variance. The first eigen vector follows similar pattern as of topography (Figure 6, 7).

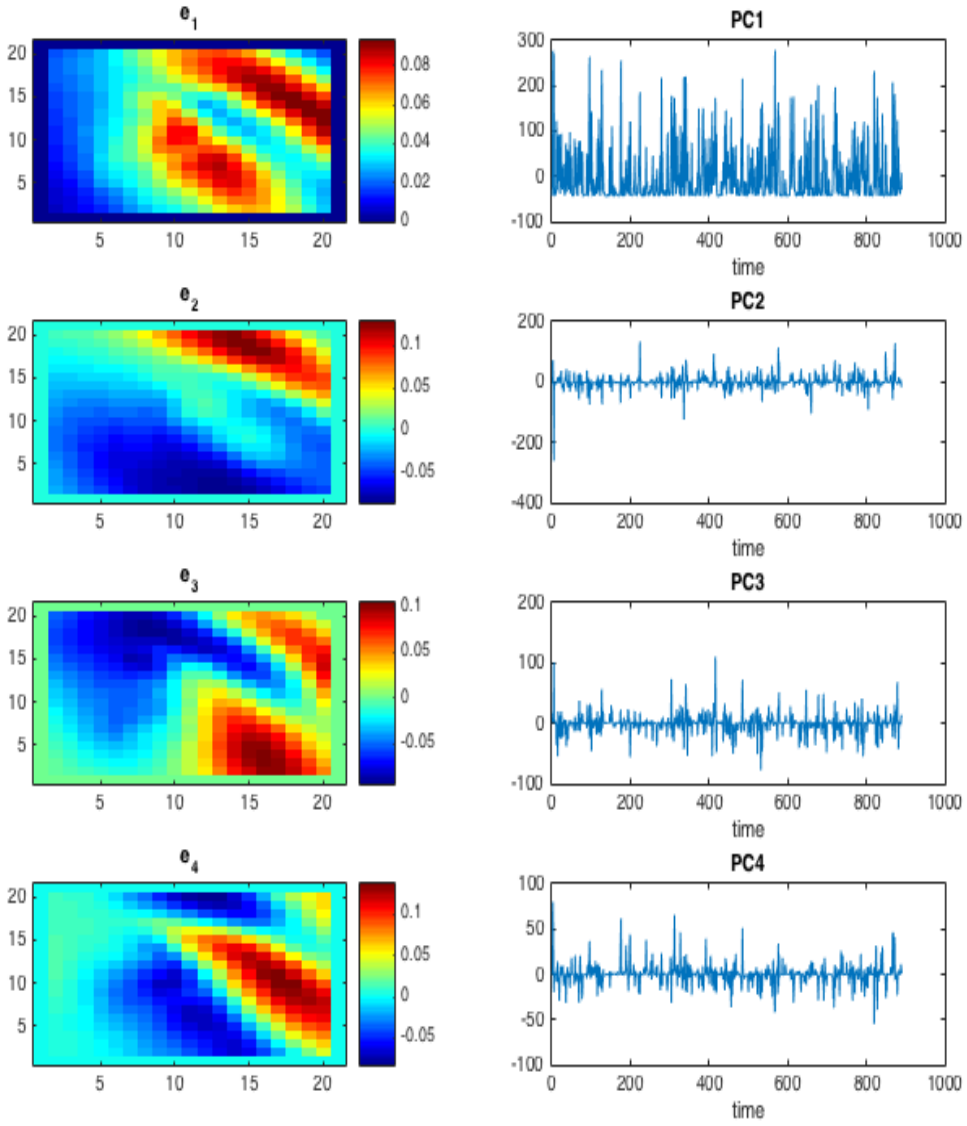
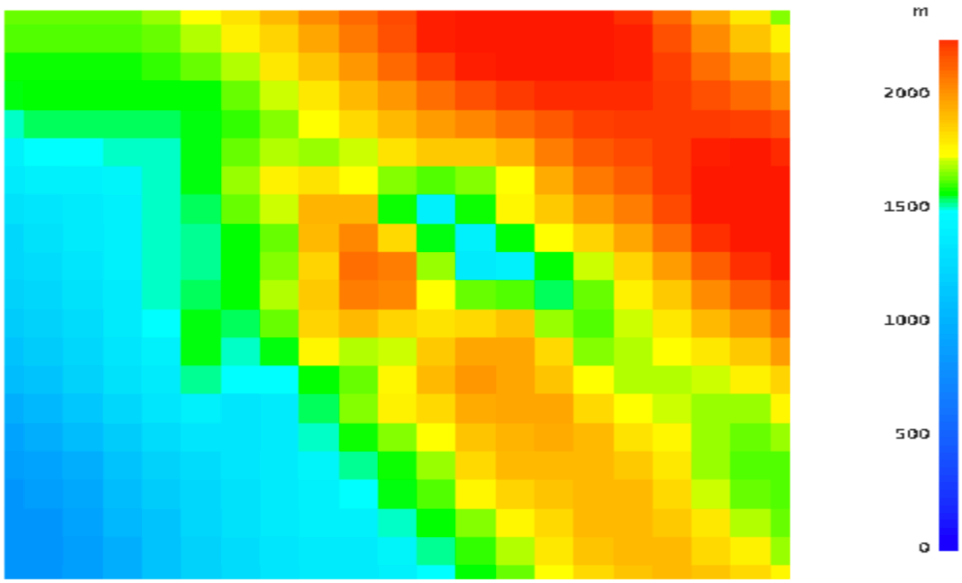


Figure 6: Significant eigen vectors and their principal components.

a)



b)

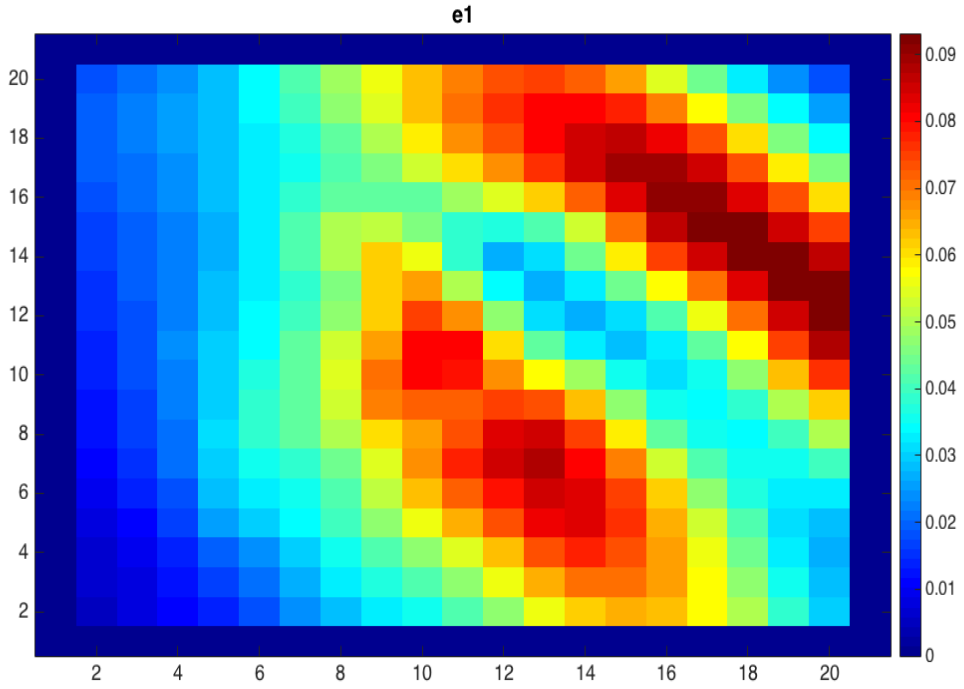


Figure 7: Comparison of a) first eigen vector and b) topography.

3.1.3 Reconstruction of daily mean precipitation

The first two PCs from WRF are used to reconstruct daily mean precipitation from the WRF and ERA-interim data. In both cases, the reconstructed data shows a strong significant correlation with the original data (RMSE = 0.180, r = 0.997) for WRF (RMSE= 1.6024, r =0.782) for ERA-interim (Figure 8).

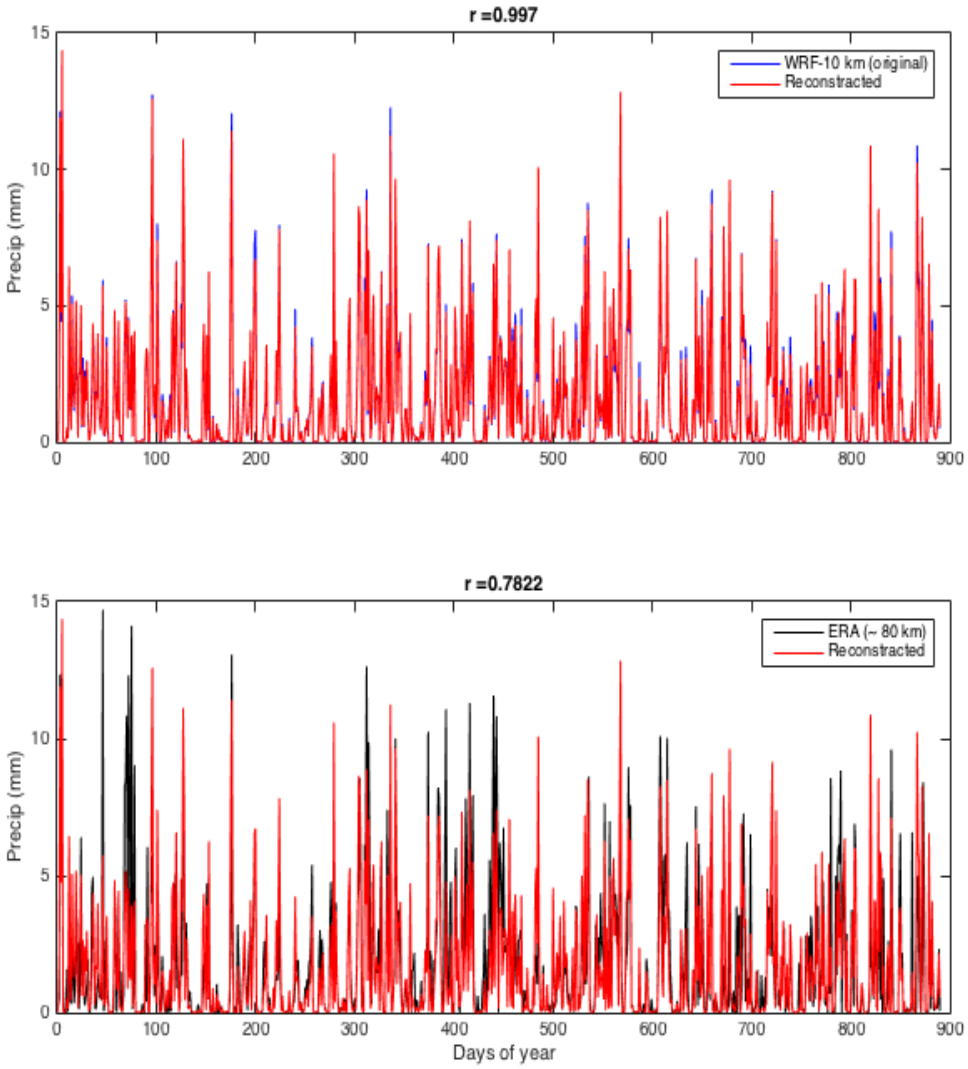


Figure 8: Reconstructed mean daily precipitation of WRF and ERA-interim data.

3.1.3 Modeling WRF-PCs from ERA-interim – PCs

Here we used three different methods such as a) Multi-layer Perceptron Neural Network (MLP NN) model, b) Multiple Linear Regression (MLR) and c) Stepwise regression to model WRF- PCs (PC1 and PC2) from five PCs obtained from the ERA interim data. According to the MLP NN method, modeled PCs shows strong significant correlation with original PCs (r = 0.78) for PC1 (Figure 9) and (r = 0.69) for PC2 (Figure 10). The result from this method is used for further analysis (i.e. reconstruction of the mean spatial patterns), which will be discussed in the next section.

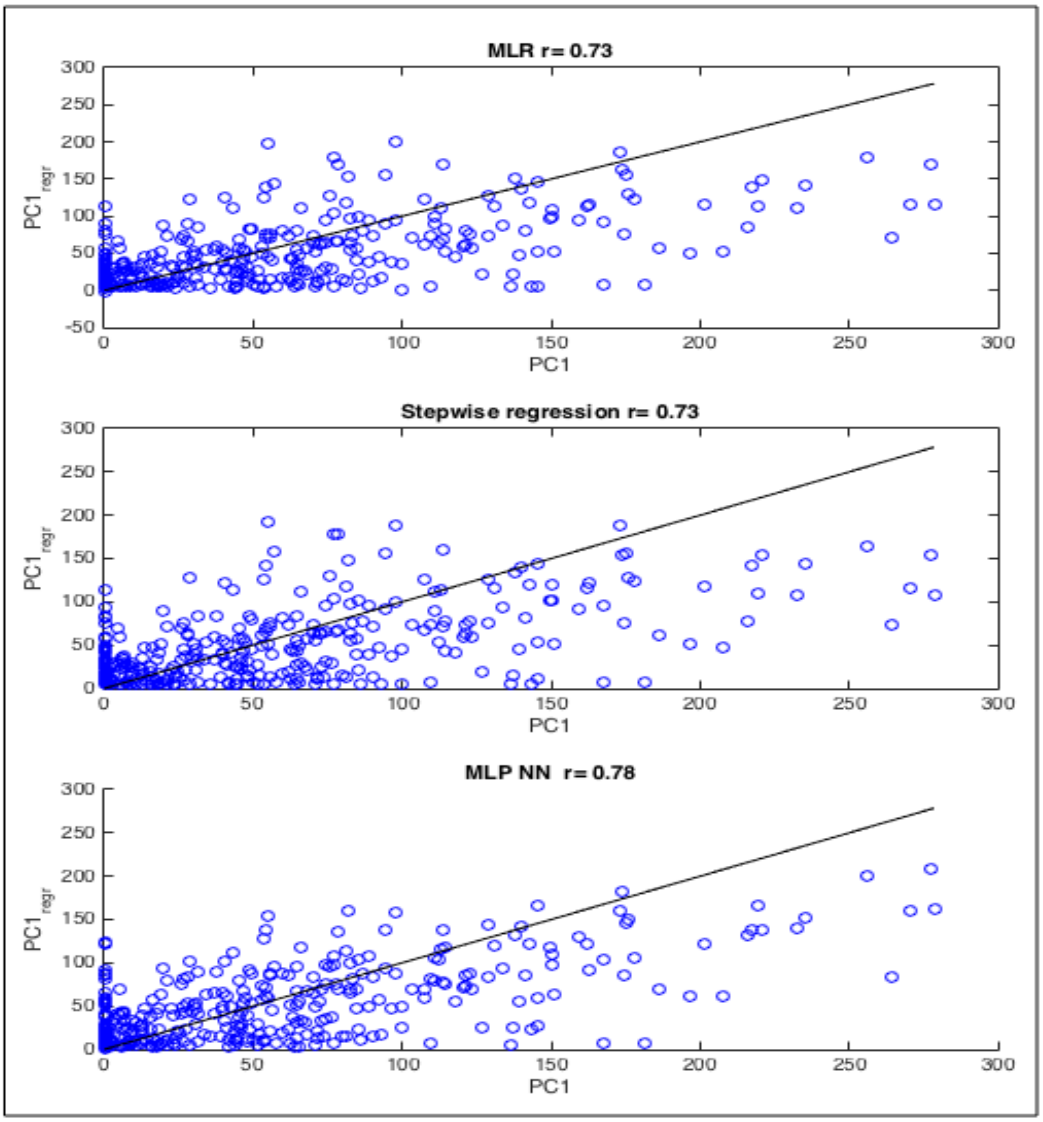


Figure 9: Modelled WRF PC1 using the first five PCs from ERA interim using a) MLR b) Stepwise regression and c) MLP NN

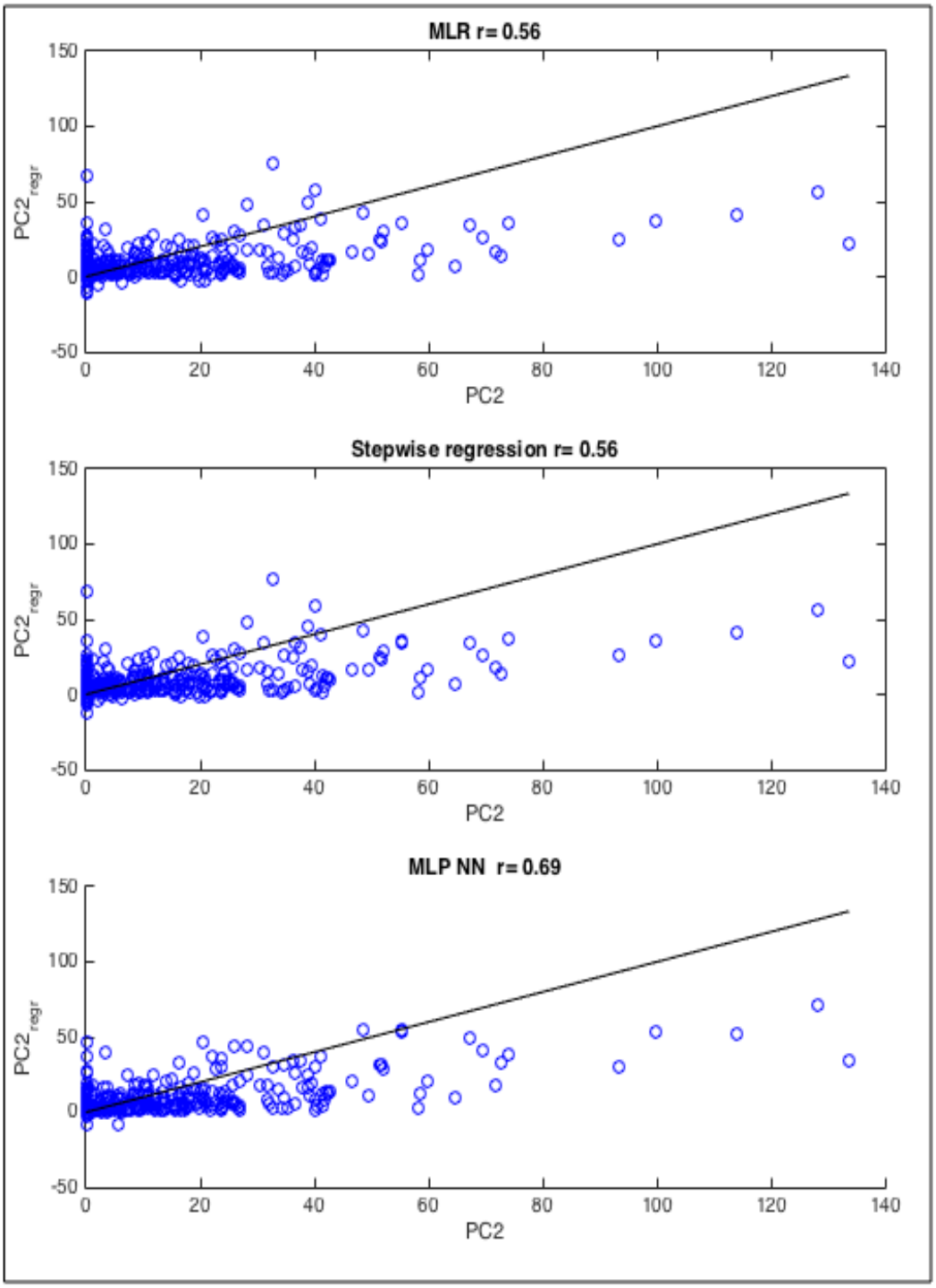


Figure 10: Modelled WRF PC2 using the first five PCs from ERA interim using a) MLR b) Stepwise regression and c) MLP NN

3.2 Results from SOM

A SOM method was used to asses most characteristic features of precipitation. Therefore, a SOM of size 7x3 (which is rescaled from 21x7) is used (Figure 11). The map shows 21 patterns and their corresponding frequency of occurrences. Node 15 is the most frequent pattern (31.9%) and the least frequent pattern is Node 7(1.6%).

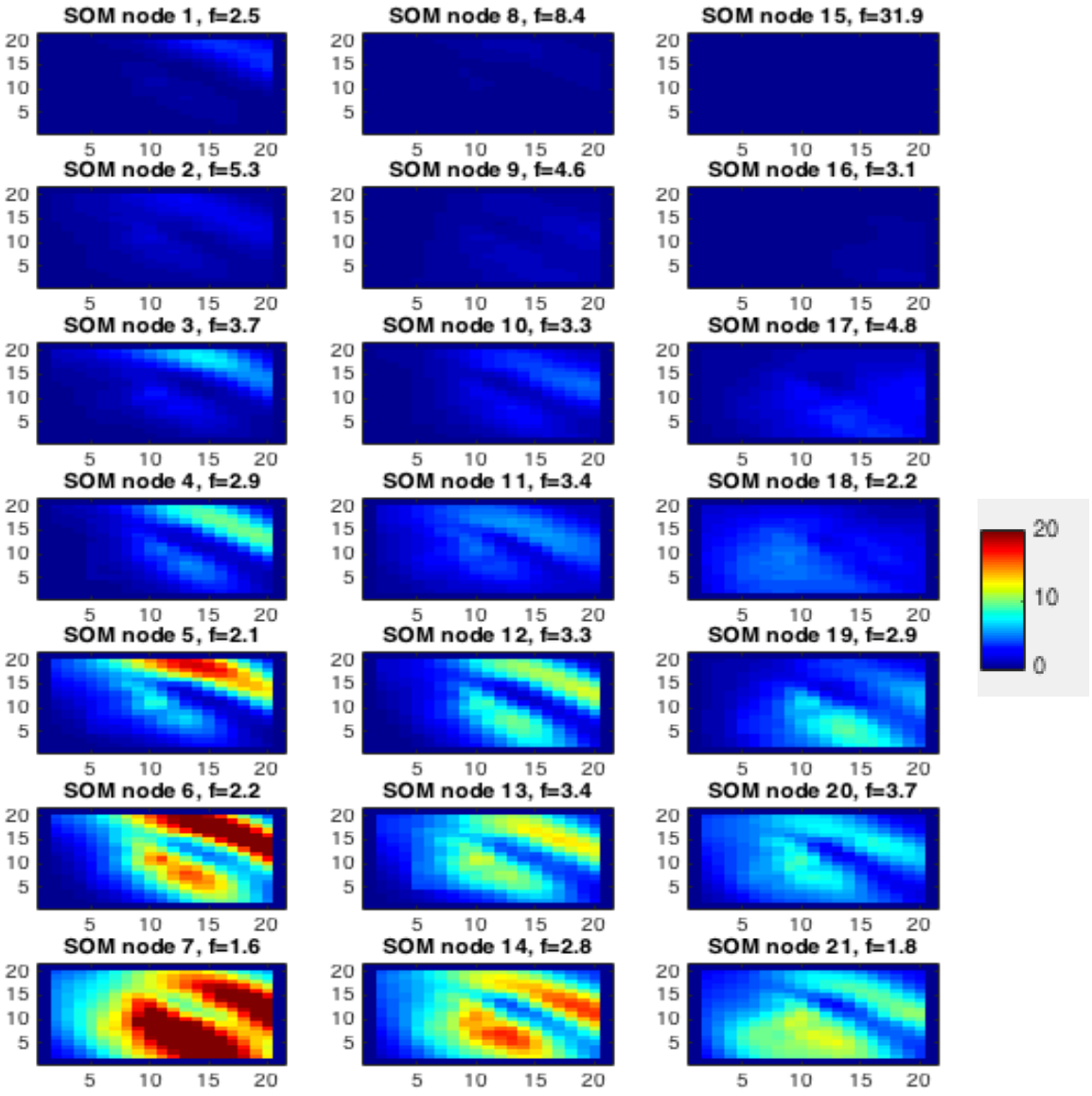


Figure 11: SOM of size 7x3 showing the frequency (percentage) of each of the patterns.

3.2.1 Dominant patterns over range of daily precipitation values

A histogram plot for the mean daily precipitation from ERA-interim (Figure 12) is used to categorize the daily precipitation values in the study period into four bins such as [0,3), [3,6), [6,9) and [9,12). Values from these bins were extracted and their corresponding dominant pattern was determined from the SOM. Each of the bins shows different dominant patterns where bin1 doesn't show a pattern following topography in comparison to the other bins. Bin 3 and 4 shows strong dependency on topography (Figure 13).

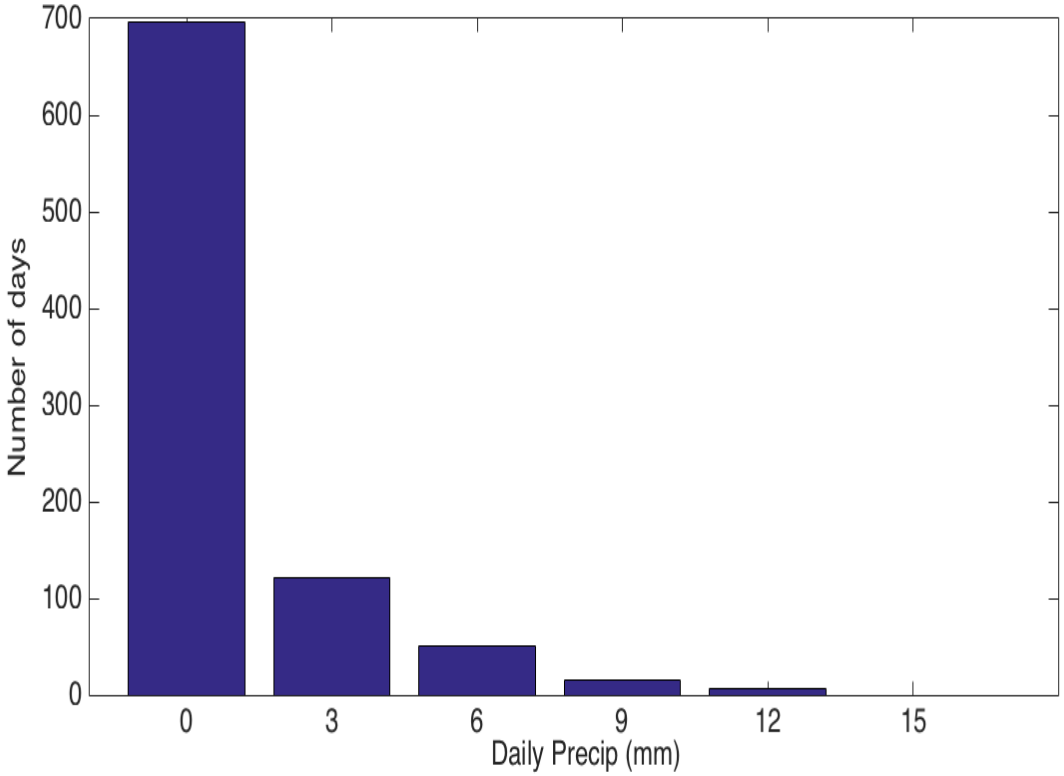


Figure 12: Mean daily precipitation from ERA-interim data.

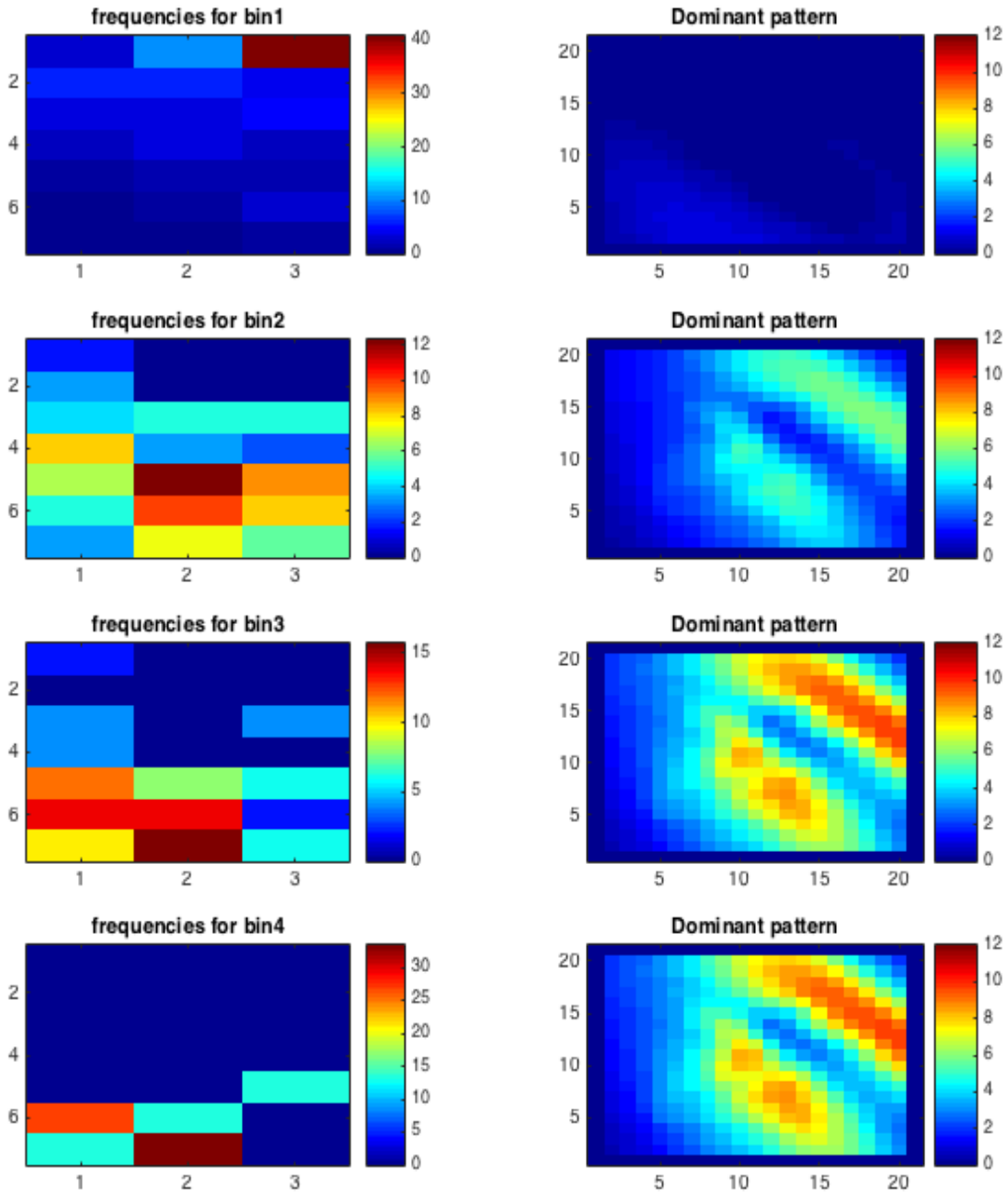


Figure 13: Dominant patterns of daily precipitation over different range of values.

3.2.2 Comparison of mean daily values (ERA-interim vs. SOM patterns)

Here we do comparison of mean daily precipitation values of days with a specific pattern from the SOM and ERA-interim data (Figure 14). Statistically significant correlation is obtained only for some of the nodes (3 (r = 04), 8 (r = 0.2), 15 (r = 0.4), 17 (r = 0.4) and 19 (r = 0.4)). Generally, ERA-interim data tend to overestimate the values in each of the nodes.

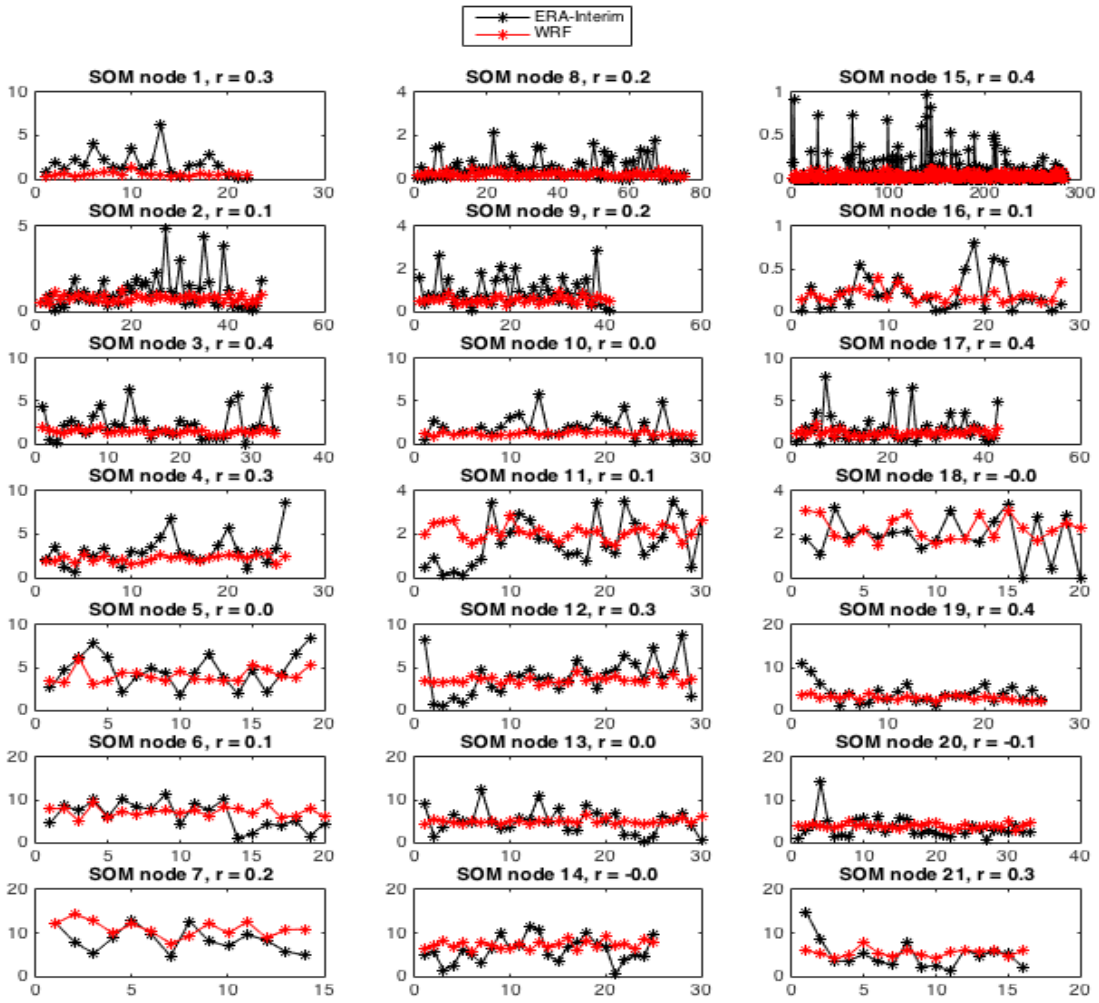


Figure 14: Mean daily precipitation from ERA-interim data vs. WRF output for all the SOM nodes.

3.2.3 Comparison of mean spatial patterns of daily precipitation

Here we use the result from section 3.1.3 to compute the mean spatial distribution of daily precipitation. Since we are interested in investigating how the PCs from WRF affect the spatial distribution of precipitation, we considered three different cases. First, we computed the mean special pattern of each of the nodes from the original SOM (Figure 15). Second, we computed mean spatial pattern of each of the nodes using the data reconstructed from WRF PCs (PC1 and PC2) (Figure 16).

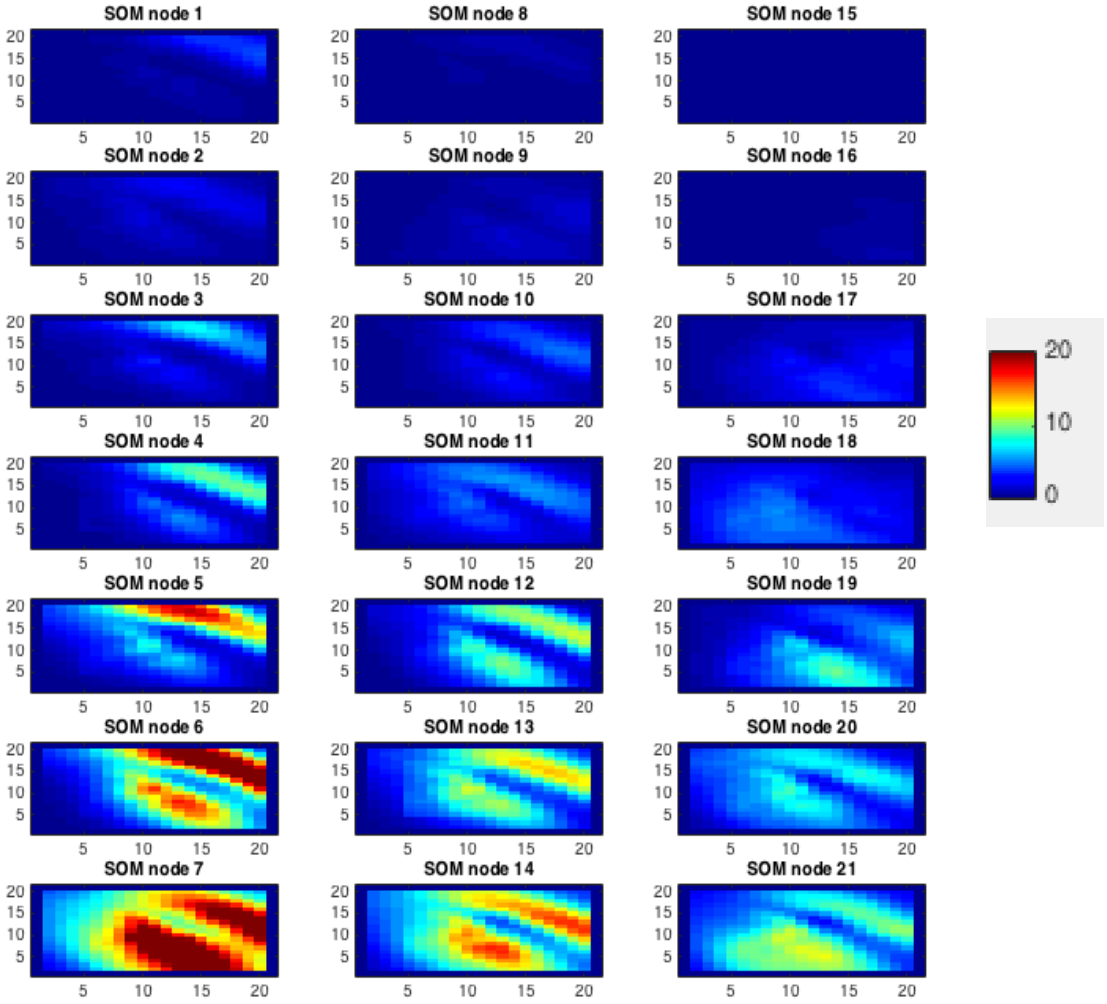


Figure 15: Mean spatial patterns for each of the nodes in the original data.

Third, we computed mean spatial patterns of each of the nodes using the reconstructed from modeled PC1 and PC2 in section 3.1.3 (Figure 17).

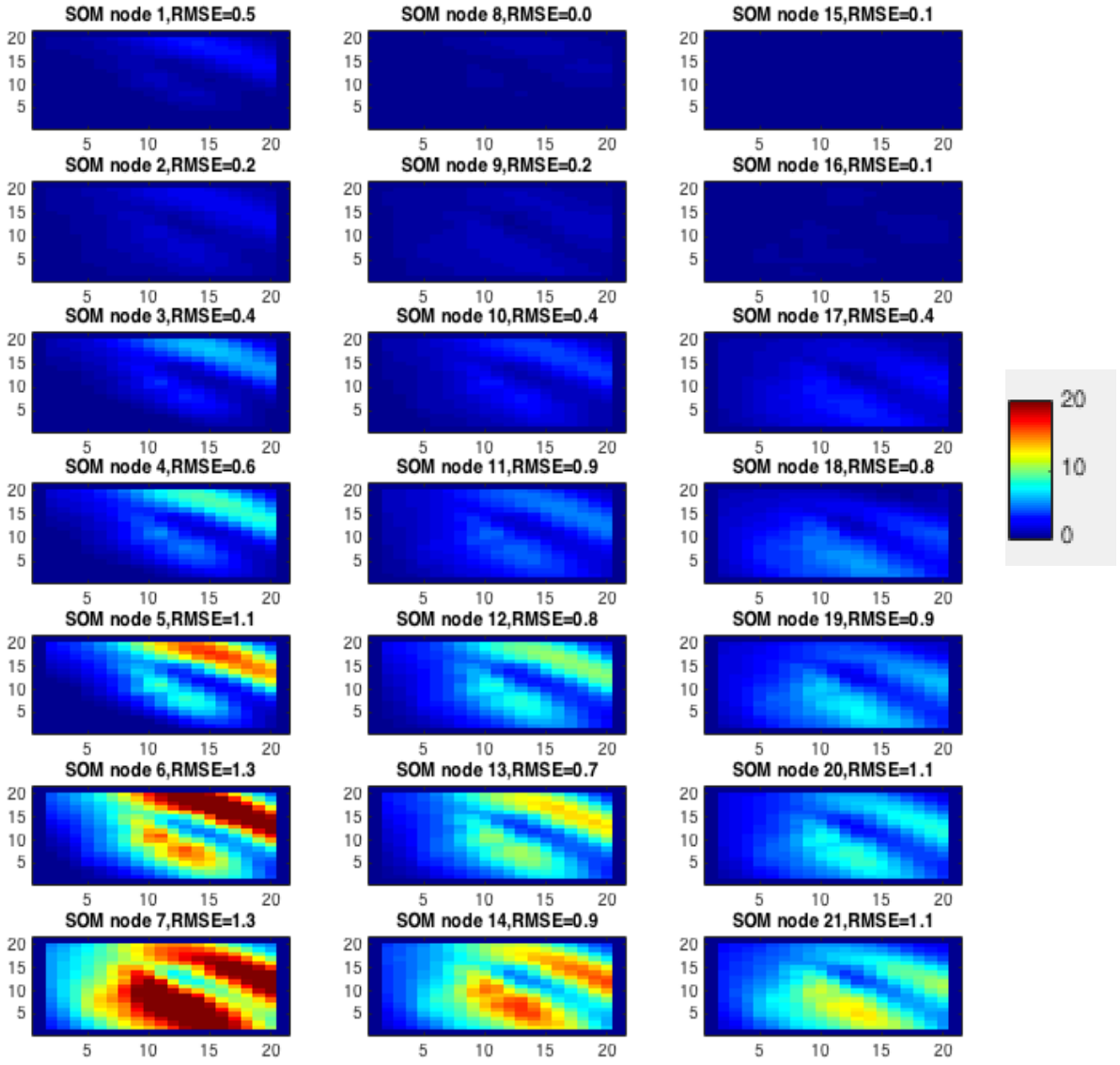


Figure 16: Mean spatial patterns of each of the nodes from a data constructed from PC1 and PC2 of WRF model output.

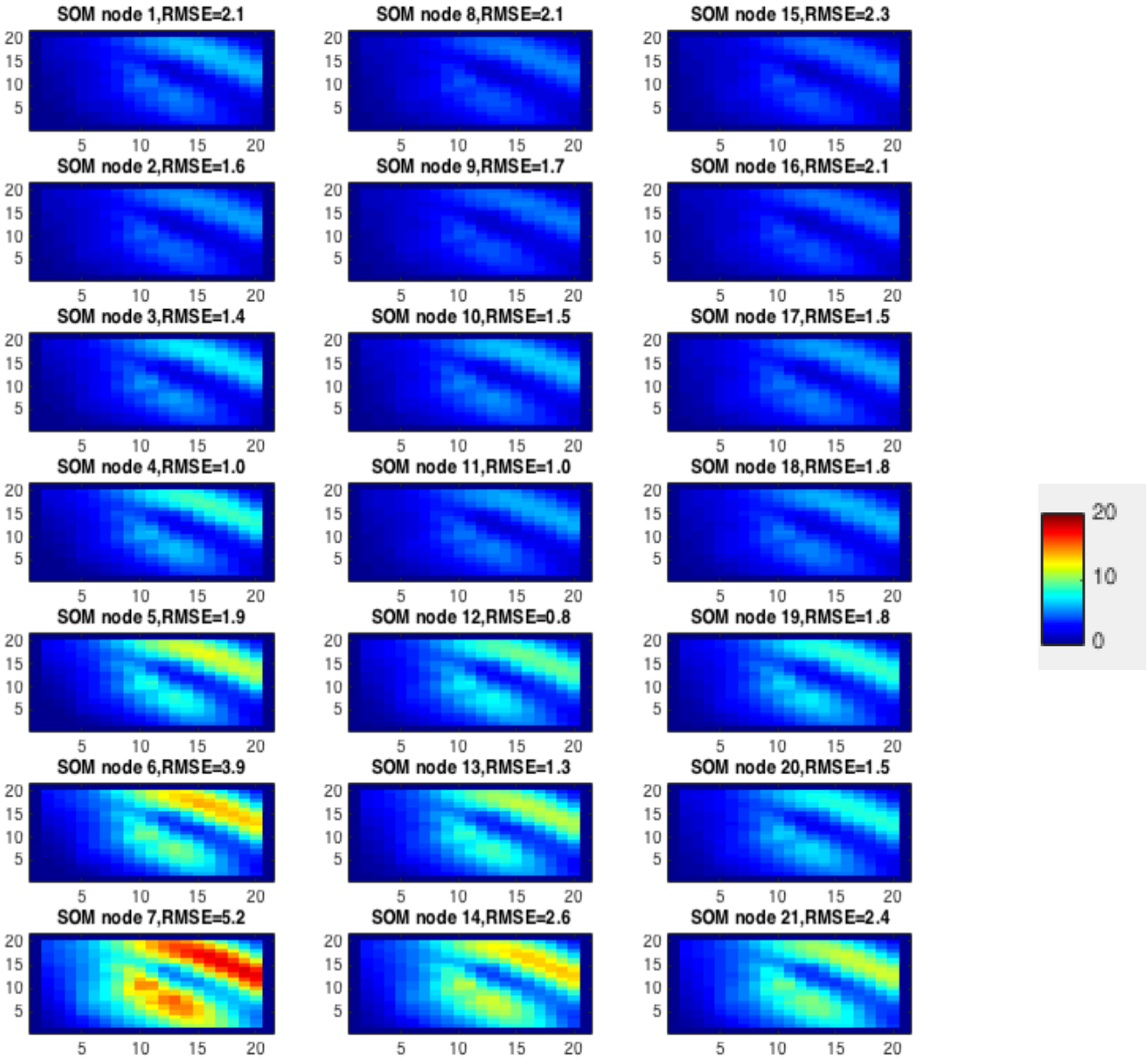


Figure 17: Mean spatial patterns for each of the nodes from a data constructed from modeled PC1 and PC2.

4. Discussion and conclusion

From the PCA analysis, the first two PCs explained 89% of the data and only the first PC explains 74% of the total variation and it follows similar pattern as topography, which might indicate that the main pattern governing precipitation in our study area is topography. Two separate MLP NN model has been used to model the first two significant WRF PCs (PC1 and PC2) using the first five ERA-interim PCs. Modeled PC1 and PC2 show strong significant correlation ($r = 0.78$, and $r = 0.69$, respectively) with the original PCs.

SOM method was applied to investigate the most characteristic features (temporal /spatial) patterns of daily precipitation. The method was applied on the downscaled data and a map of 7x3 (rescaled from the suggested map size of 21x6). The SOM shows that the most frequently occurring pattern (31.9%) is the one with lower precipitation, which is distributed throughout the domain not specifically, following the topography. The least occurring pattern (1.6%) is the one with higher precipitation showing higher values of daily precipitation over higher latitudes and vise versa. Similarly, the dominant pattern of daily precipitation over a range of lower values [0 – 3mm) is similar to the most frequent pattern (node 15) and precipitation over the range of higher values [6 – 9mm) and [9 – 12 mm) follows a distinct pattern following topography.

We also performed comparison of mean spatial patterns of each of the nodes using a) the original data b) data reconstructed from original PCs from WRF (PC1 and PC2) and c) data reconstructed from modeled PCs. The result shows that the mean spatial patterns in case (b) show a lower RMSE than case (c) for each of the nodes. However, the distribution of the error seems similar in such away that higher RMSE goes to nodes with higher precipitation values and vise versa. Only few nodes show statistically significant correlation when comparing mean daily precipitation of days of specific patterns with same days from ERA-interim data and generally ERA-interim data overestimated the values in each of the nodes.

In conclusion, the results from both PCA and SOM provide some insight on the temporal and spatial distribution of precipitation. The dynamical downscaling with WRF also enhances understanding of local scale variability and patterns of daily precipitation over the study area. The overall result shows that there is higher uncertainty in predicting long term higher values (extreme events) of precipitation using the modeled PCs. For future work the methods (PCA and SOM) might work better when using a larger data set (e.g. 20 years) and the result from this project should be validated using different scenarios.

5. References

1) Mills G (1995) Principal Component Analysis of Precipitation and Rainfall Regionalization in Spain. Theoretical and Applied Climatology, 50 (169 - 183)

2) Stathis D and Myronidis D (2009) Principal component analysis of precipitation in Thessaly region (Central Greece). Global NEST Journal, 11 (467 - 476).

3) Othman M, Ashaari Z and Mohamad N (2015) Long-term daily rainfall pattern recognition: Application of principal component analysis. Procedia Environmental Sciences 30 (127 - 132).