Valliappa Lakshmanan · Eric Gilleland
Amy McGovern · Martin Tingley *Editors*

Machine Learning and Data Mining Approaches to Climate Science

Proceedings of the 4th International Workshop on Climate Informatics

Springer

Valliappa Lakshmanan • Eric Gilleland
Amy McGovern • Martin Tingley
Editors

Machine Learning and Data Mining Approaches to Climate Science

Proceedings of the 4th International Workshop on Climate Informatics

Springer

Chapter 3
Teleconnections in Climate Networks: A Network-of-Networks Approach to Investigate the Influence of Sea Surface Temperature Variability on Monsoon Systems

Aljoscha Rheinwalt, Bedartha Goswami, Niklas Boers, Jobst Heitzig, Norbert Marwan, R. Krishnan, and Jürgen Kurths

Abstract We analyze large-scale interdependencies between sea surface temperature (SST) and rainfall variability. We propose a novel climate network construction scheme which we call *teleconnection climate networks* (TCN). On account of this analysis, gridded SST and rainfall data sets are coarse grained by merging grid points that are dynamically similar to each other. The resulting

A. Rheinwalt (✉)
Potsdam Institute for Climate Impact Research, Potsdam, Germany

Humboldt-Universität zu Berlin, Berlin, Germany

University of Potsdam, Potsdam, Germany
e-mail: aljoscha@pik-potsdam.de

B. Goswami
Potsdam Institute for Climate Impact Research, Potsdam, Germany

University of Potsdam, Potsdam, Germany

N. Boers
Potsdam Institute for Climate Impact Research, Potsdam, Germany

Department of Physics, Humboldt University, Berlin, Germany
e-mail: boers@pik-potsdam.de

J. Heitzig • N. Marwan
Potsdam Institute for Climate Impact Research, Potsdam, Germany

R. Krishnan
Indian Institute of Tropical Meteorology, Pune, India

J. Kurths
Potsdam Institute for Climate Impact Research, Potsdam, Germany

Department of Physics, Humboldt University, Berlin, Germany

Institute for Complex Systems and Mathematical Biology, University of Aberdeen, Aberdeen, UK

Department of Control Theory, Nizhny Novgorod State University, Nizhny Novgorod, Russia
e-mail: kurths@pik-potsdam.de

V. Lakshmanan et al. (eds.), *Machine Learning and Data Mining Approaches to Climate Science*, DOI 10.1007/978-3-319-17220-0_3

clusters of time series are taken as the nodes of the TCN. The SST and rainfall systems are investigated as two separate climate networks, and teleconnections within the individual climate networks are studied with special focus on dipolar patterns. Our analysis reveals a pronounced rainfall dipole between Southeast Asia and the Afghanistan-Pakistan region, and we discuss the influences of Pacific SST anomalies on this dipole.

Keywords Clustering • Precipitation dipole • Teleconnections • Complex networks • Time series analysis

3.1 Introduction

Precipitation on the Asian continent is known to be influenced by large-scale atmospheric processes like the Hadley and Walker circulation. However, the intricate interplay of different atmospheric processes and how they influence precipitation variability are still not completely understood. Here, we study long-range interrelations within the precipitation system as well as between precipitation and sea surface temperature (SST) dynamics. Our aim is to shed light on the spatial structure of such teleconnections, with a special focus on precipitation dipoles and how they are influenced by SST variability.

For this purpose, we employ the climate network approach by representing the interrelations between climatic time series as complex networks (Boers et al. 2013, 2014; Donges et al. 2009a,b; Ebert-Uphoff and Deng 2012; Malik et al. 2012; Tsonis and Roebber 2004; Tsonis et al. 2006; Yamasaki et al. 2008). The SST and the precipitation system are studied as two separate networks and the interrelations between them by their cross topology.

So far, empirical orthogonal functions (EOFs), which are derived from principal component analysis of covariance matrices, are commonly used for a spatial analysis of teleconnections in climatological data (Ghil et al. 2002). While certainly very useful in many situations, they carry certain caveats in such analyses: First, if the data are not normally distributed, the corresponding EOFs will in general, while uncorrelated, not be statistically independent (Monahan et al. 2009). Second, even if they are independent, EOFs do not necessarily uniquely correspond to climatological mechanisms (Dommenget and Latif 2002). Third, and maybe most importantly, analyses based on the covariance matrix will only be able to capture linear dependencies. This might be considered insufficient in view of the strong nonlinearities involved in climatic interactions. Climate network can be considered as a complementary approach to study spatial patterns of climatic interrelations, which do not suffer from these statistical problems if derived from a nonlinear similarity measure. Furthermore, since teleconnections are not directly represented as links in EOFs, they have to be deduced from the spatial patterns. Although this might be possible for simple teleconnection structures, it becomes challenging for more complicated ones.

Nonetheless, the common way of climate network construction is not suitable for the investigation of teleconnections as well. There, traditionally a pairwise similarity

analysis between all pairs of time series is performed, for instance, by the use of Pearson's correlation coefficient (Donges et al. 2009b; Tsonis et al. 2006). However, climate networks are spatially embedded networks, and the similarity between time series is strongly dependent on their spatial distance (Rheinwalt et al. 2012): Two time series that are spatially close to each other are likely to be more similar than two time series which are far away from each other in space. By focusing only on strong similarities as in most climate network studies, networks have essentially only short links, which led to the investigation of paths in climate networks (Donges et al. 2009a).

Here we propose an approach that groups all time series by similarity into clusters. A related idea was also pursued in Hlinka et al. (2014). We use a specific clustering scheme that typically provides spatially connected clusters due to the distance dependence of the similarities in climate systems. In other words, these clusters are localized regions of high resemblance according to the dynamics of the corresponding time series. Each cluster will in our approach be represented by a single time series, and only the similarity structure between these representatives will be explored. By doing so we do not only reduce the dimensionality of the network, but we more importantly constructed a climate network that is reduced to its teleconnections. We will refer to these networks as *teleconnection climate networks* (TCN).

3.2 Method

In order to group time series by similarity, we use the standard fast greedy hierarchical agglomerative *complete linkage clustering* (Defays 1977). This clustering is performed in a metric space with dissimilarities between time series as distances. In this study we focus on the Spearman's rho correlation coefficient as the similarity measure in order to capture not only linear but also other monotonic relationships and in order to avoid problems of skewed distributions in precipitation data. In our case of standardized anomalies that have zero mean and unit variance, this coefficient is proportional to the dot product between the ranked variables and can be interpreted as the cosine of the angle θ between these two ranked variables. More precisely, the Spearman's rho $\varrho_{X,Y}$ between two ranked time series X and Y is given by

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} \equiv \frac{X \cdot Y}{\|X\| \, \|Y\|} = \cos(\theta_{X,Y}) \,. \tag{3.1}$$

This angle θ in radians between two time series is a distance that we use as the dissimilarity measure for the clustering.

Statistical significance of Spearman's rho values is estimated using twin surrogates.[1] These carry the advantage of preserving dynamical features of the original

[1] Due to the short length of time series we obtain the twin surrogates without embedding.

time series in contrast to bootstrapping methods (Marwan et al. 2007; Romano et al. 2009; Thiel et al. 2006, 2008). For each pair of time series, we test against the null hypothesis that they are independent realizations of the same dynamical system. Upon repeating this for all pairs of time series, we pick the maximum threshold corresponding to the 98 % confidence level as a global significance threshold $T^{0.98}(\varrho)$.

We intend to group time series into clusters in such a way that all correlation values between time series within a given cluster are statistically significant. This is achieved by the use of the *complete linkage clustering* scheme that is also known as *farthest neighbor clustering*. The distance measure between two clusters U and V is in this scheme defined as

$$D(U,V) = \max_{X\in U, Y\in V} d(X,Y) = \max_{X\in U, Y\in V} \theta_{X,Y}\,. \tag{3.2}$$

We cut the resulting dendrogram at the distance d_{crit} that corresponds to the significance threshold of all pairwise correlation values, i.e., $d_{\text{crit}} = \arccos(T^{0.98}(\rho))$. This yields the maximum number of partitions of the set of time series such that for any two clusters U and V holds, $D(U,V) \geq d_{\text{crit}}$, which is the same as the minimum number of partitions such that for any two time series $X, Y \in U$ in any given cluster U, we have $\theta_{X,Y} < d_{\text{crit}}$. This clustering method does not only assure that all time series within a cluster are significantly correlated when cutting the dendrogram at d_{crit} but also avoids the *chaining phenomenon* of the *single linkage clustering* where a set of time series might form a cluster although only a few time series are actually close to each other (Everitt et al. 2001). The clustering reduces the dimensionality of the problem by merging dynamically similar time series into clusters, which will serve as nodes for the *teleconnection climate networks* (TCN) that will be constructed in the following.

More specifically, a TCN node is represented by a single time series from the corresponding cluster. Although there are clustering schemes, such as the *k-means clustering* (MacQueen et al. 1967), that suggest a certain member of a cluster as a representative, the in this study anticipated *complete linkage clustering* does not. Also, since cluster sizes vary, special care has to be taken when choosing a representative time series for a cluster. For instance, the point-wise mean of all time series within a cluster would be influenced by the size of the cluster. Instead we pick the time series with the highest average correlation to all other time series within that cluster as a representative for that cluster. This also has the advantage that the representative time series retain the original variabilities.

The TCN is now constructed by computing ϱ for all pairs of representative time series and assigning the corresponding values as link weights. We remove all links from the TCN that have a weight equal or below $T^{0.98}(\varrho)$.

We note that TCN could as well be studied using node-weighted network measures (Heitzig et al. 2012; Wiedermann et al. 2013). Although not a focus of this study, this is an interesting topic of future research.

3.3 Application

We apply the proposed methodology to precipitation data for the Asian continent together with a global SST data set. We will in the following investigate dipole structures in the precipitation system and how these dipoles are influenced by SST variability.

3.3.1 Data

We use monthly time series for the years 1982–2008: SST data is obtained from the NOAA Optimum Interpolation SST V2 on a one-by-one-degree grid (Reynolds et al. 2002), and precipitation data over land is taken from the APHRODITE V1101 daily precipitation data product on a 0.25×0.25 degree grid (Yatagai et al. 2012). In the latter data set, monthly mean values were calculated from daily values in a preprocessing step. We study monthly anomalies, in contrast to the monthly mean values itself, where the seasonal cycle would dominate correlation coefficients. Anomalies are calculated by subtracting from each value the long-term mean for that month and dividing by the corresponding long-term standard deviation.

3.3.2 Coarse Graining

Based on the significance tests explained above, we obtain significance thresholds $T^{0.98}(\varrho) = 0.199$ for the precipitation data set and $T^{0.98}(\varrho) = 0.494$ for the SST data set. Hence, we cut the Asian precipitation dendrogram at $\varrho = 0.2$. This leads to 111 precipitation clusters which are shown in Fig. 3.1. The geographical location of representative time series is depicted as black dots. With an initial number of 31624 time series, the coarse graining reduces the number of time series by a factor of $\approx$ 285. While the minimum correlation within clusters is 0.2, the average correlation within a cluster has a much higher value of 0.7.

We cut the global SST dendrogram at a threshold of $\varrho = 0.5$. This leads to 1419 SST clusters as shown in Fig. 3.2. With an initial number of 40780 SST time series, the coarse graining reduces the number of time series only by a factor of $\approx$29. This lower reduction is due to the relatively coarser spatial resolution of the SST data set. The correlation coefficient between SST time series within a cluster is, with an average value of 0.8, even higher than for the precipitation clustering.

3.3.3 Dipoles

In order to focus on precipitation dipoles, we reduce the precipitation TCN by removing all nodes that do not even have a single significant link with a negative link

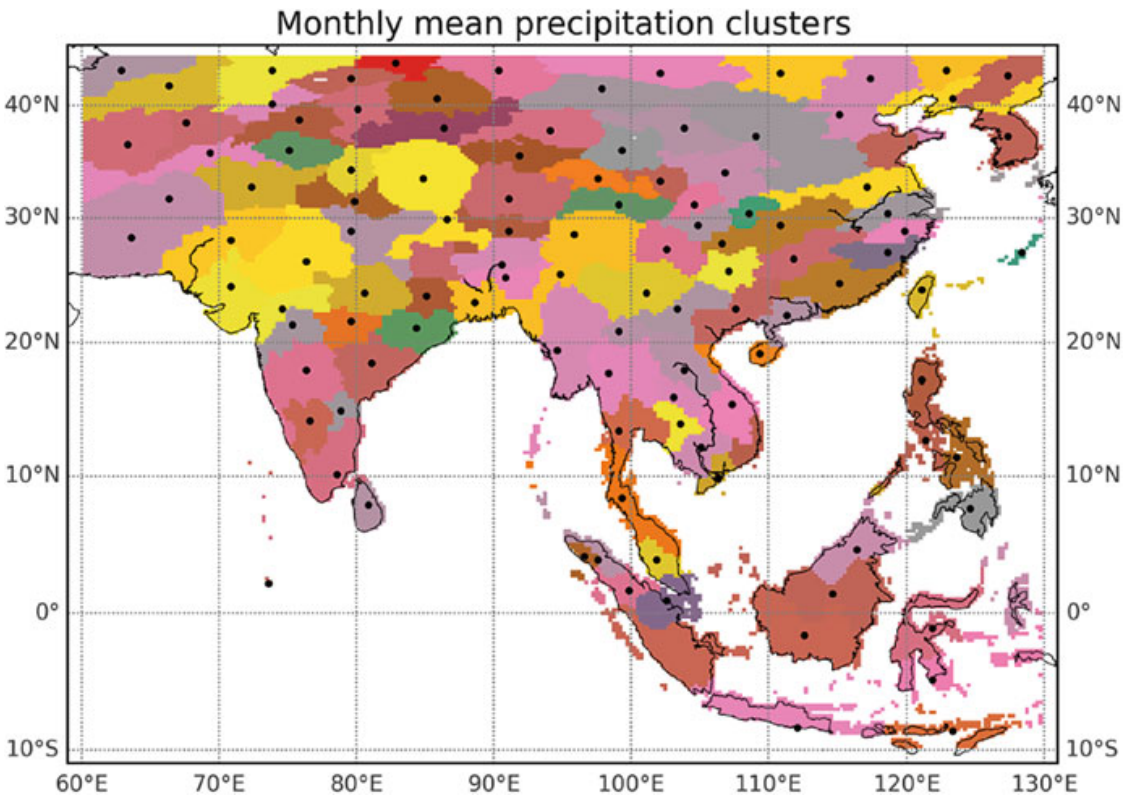

Fig. 3.1 Clustering of the precipitation data using the arccosine of the Spearman's rank correlation as a distance metric. All time series within a cluster are significantly correlated to each other. This corresponds to a minimum correlation of 0.2 between time series within a cluster. However, the average correlation within a cluster is on average 0.7. Geographical locations of representative time series for clusters are depicted as *black dots*

weight. Note that we understand dipoles as anticorrelations between representative time series. The resulting network reflects the dipole structure that is captured from the APHRODITE data set for the considered time period. It consists of only 36 anticorrelation links (red) and 83 correlation links (blue) (see Fig. 3.3).

3.3.4 Networks of Climate Networks

Given the two sets of representative time series for the precipitation data set as well as for the SST data set, we estimate all pairwise lagged correlation coefficients between these two sets. We consider possibly lagged correlation, because teleconnections between Asian precipitation and the global SST field can in general occur with a delay even on monthly scales. We employ a simple maximum correlation approach as follows. We focus on the influence of SST variability on precipitation and thus only consider lags that correspond to SST dynamics preceding precipitation dynamics, where we consider only lags up to 12 month. As link weights we take the first local maximum of Spearman's rho over this range of lags. A similar approach was taken, for example, in Yamasaki et al. (2008).

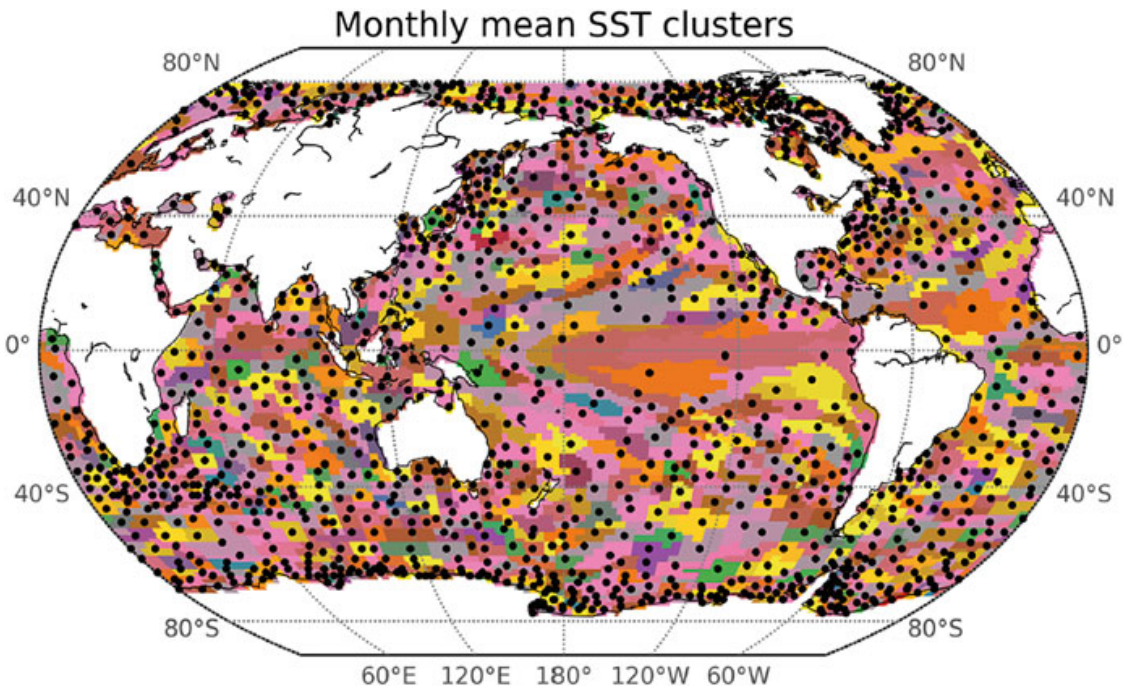

Fig. 3.2 Clustering of the SST data using the arccosine of the Spearman's rank correlation as a distance metric. All time series within a cluster are significantly correlated to each other, which corresponds to a minimum correlation of 0.5 between time series within a cluster. The average correlation within a cluster is on average 0.8. Geographical locations of representative time series for clusters are depicted as *black dots*

In order to understand the influence of SST variability on the obtained Asian precipitation dipole, we examine cross-links of nodes from the Southeast Asian pole (see Fig. 3.3). All the nodes in this region, marked as yellow dots in Fig. 3.4, experience a spatially very similar influence from the SST network (not shown). Thus, we show the mean correlation from the SST network to these precipitation nodes (see Fig. 3.4).

3.4 Results and Discussion

Using the proposed method of TCN construction, we find a strikingly pronounced precipitation dipole between the Southeast Asian region and the Afghanistan-Pakistan region. This dipole has, for example, been described in Barlow et al. (2005). In that study, the authors partly explain its occurrence by an interplay of the Madden-Julian oscillation and the African-Arabian jet stream. Furthermore, this dipolar pattern is most likely related to the lateral component of the Asian monsoon system (Trenberth et al. 2000; Webster et al. 1998, 1999).

The Southeast Asian region, in the precipitation network represented by nodes marked as yellow dots in Fig. 3.4, is a major deep convection area of the considered precipitation network. Convection is forced by solar heating and forms a rising

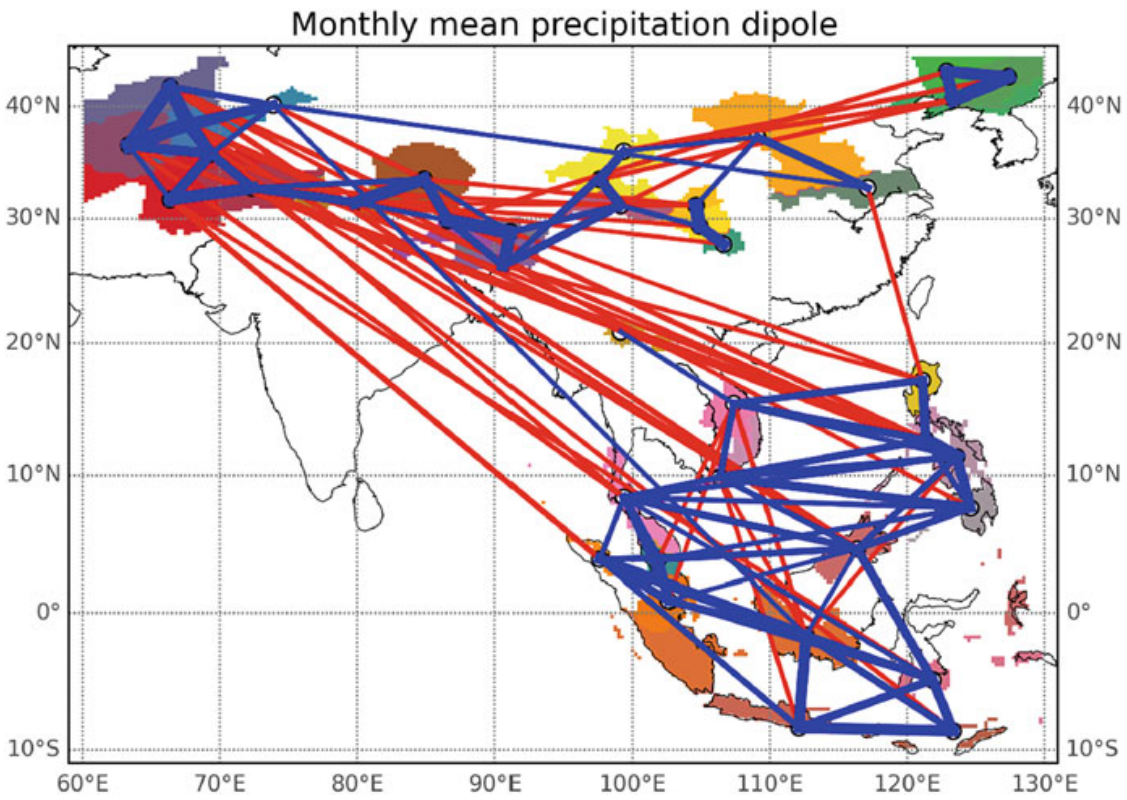

Fig. 3.3 The precipitation TCN reduced to nodes that have significant anticorrelations (*red links*) and correlations (*blue links*) to other representative precipitation time series. Link thickness is proportional to absolute link weight. Links are drawn between geographical positions of representative time series, and the corresponding clusters are colored. Observe the pronounced precipitation dipole between Southeast Asia and the Afghanistan-Pakistan region

branch of the Hadley cell in this area but is also modulated by the Walker circulation (Gill 1980). This modulating effect explains the negative correlation values between precipitation in the Southeast Asian region and SST anomalies in the eastern central tropical Pacific observed in Fig. 3.4: The Walker circulation causes upward atmospheric motion at the western boundary of the tropical Pacific and downward motion at the eastern boundary. If the Walker circulation weakens as under El Niño conditions, convection is suppressed in the Southeast Asian region, resulting in reduced precipitation. At the same time, upwelling of cold water in the eastern Pacific ocean is reduced, which causes positive SST anomalies in the eastern and central tropical Pacific. Correspondingly, a strengthened Walker circulation causes stronger convection in the Southeast Asian region and negative SST anomalies in the eastern and central tropical Pacific.

On the other hand, we also observe a V-shaped pattern of positive correlation values in Fig. 3.4, with two branches extending to the subtropics. These two branches follow the climatological orientation of the trade winds in this region, and we suggest the following explanation for this pattern: Since the specific humidity of the low-level atmosphere rises with temperature, and the air temperature is in turn coupled to the SSTs, air parcels arriving at the Southeast Asian region

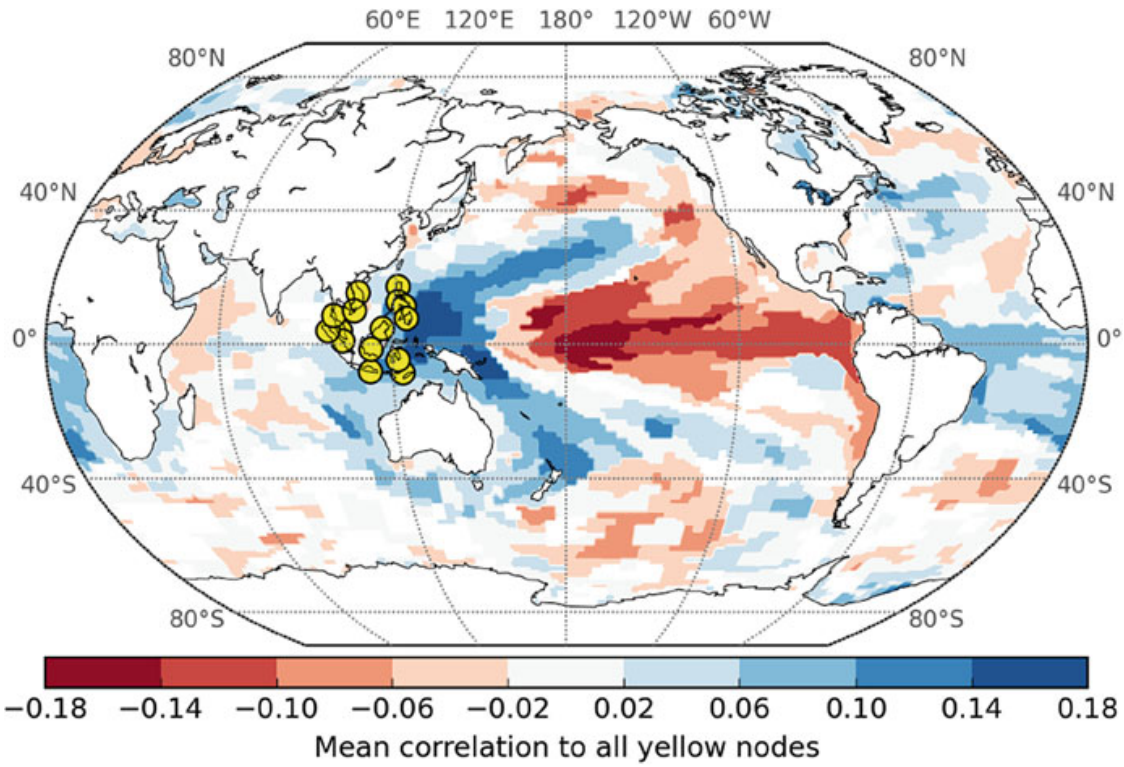

Fig. 3.4 Mean correlation between monthly precipitation anomalies in the Southeast Asian pole of the dipole (*yellow dots*) to the global SST field. Observe the negative (*red*) mean correlation values between this pole and the SSTs in the tropical central and eastern Pacific, as well as the positive (*blue*) mean correlation pattern extending from the pole to the subtropics

will carry the more (less) moisture the warmer (cooler) the SSTs are along the trajectory of the trade winds from the subtropics. This modulates the water vapor content of the air that rises in the Southeast Asian region due to the convection discussed in the last paragraph and hence the amount of precipitation. We note that this mechanism should also apply to the tropical Pacific, but there, its influence is strongly overprinted by the Walker circulation.

3.5 Conclusion

We proposed a new framework to construct multivariate climate networks from observational data. This framework is designed to study long-range interrelations, i.e., teleconnections, by first merging dynamically similar time series into clusters and then investigating connections between these clusters. We applied our approach to SST data as well as precipitation data over the Asian continent and coupled the two separate networks obtained for each variable to a network of climate networks in order to study the impacts of SST variability on teleconnections in the precipitation network. Our analysis reveals a pronounced precipitation dipole between Southeast Asia and the Afghanistan-Pakistan region, which may be controlled by an interplay

of the Madden-Julian oscillation, and the African-Arabian jet stream. Results obtained from the coupled network-of-networks analysis further suggest that trade winds from the subtropics as well as the Walker circulation over the tropical Pacific in turn modulate this dipole.

Acknowledgements Funded by DFG, project *Investigation of past and present climate dynamics and its stability by means of a spatio-temporal analysis of climate data using complex networks* (MA 4759/4-1). Further support by DFG/FAPESP IRTG 1740/TRP 2011/50151-0.

References

Barlow M, Wheeler M, Lyon B, Cullen H (2005) Modulation of daily precipitation over southwest Asia by the Madden-Julian oscillation. Mon Weather Rev 133(12):3579–3594

Boers N, Bookhagen B, Marwan N, Kurths J, Marengo J (2013) Complex networks identify spatial patterns of extreme rainfall events of the South American Monsoon System. Geophys Res Lett 40(16):4386–4392. Wiley Online Library

Boers N, Bookhagen B, Barbosa HMJ, Marwan N, Kurths J, Marengo JA (2014) Prediction of extreme floods in the eastern Central Andes based on a complex networks approach. Nat Commun 5:5199. Nature Publishing Group

Defays D (1977) An efficient algorithm for a complete link method. Comput J 20(4):364–366. Br Computer Soc

Dommenget D, Latif M (2002) A cautionary note on the interpretation of EOFs. J Clim 15(2):216–225

Donges JF, Zou Y, Marwan, N, Kurths J (2009a) The backbone of the climate network. EPL (Europhys Lett) 87(4):48007. IOP Publishing

Donges JF, Zou Y, Marwan, N, Kurths J (2009b) Complex networks in climate dynamics. Eur Phys J Spec Top 174(1):157–179. Springer

Ebert-Uphoff I, Deng Y (2012) Causal discovery for climate research using graphical models. J Clim 25(17):5648–5665

Everitt BS, Landau S, Leese M (2001) Cluster analysis. Arnold, London

Gill A (1980) Some simple solutions for heat-induced tropical circulation. Q J R Meteorol Soc 106(449):447–462. Wiley Online Library

Ghil M, Allen MR, Dettinger MD, Ide K, Kondrashov D, Mann ME, Robertson AW, Saunders A, Tian Y, Varadi Fet al (2002) Advanced spectral methods for climatic time series. Rev Geophys 40(1):3–1

Heitzig J, Donges JF, Zou Y, Marwan N, Kurths J (2012) Node-weighted measures for complex networks with spatially embedded, sampled, or differently sized nodes. Eur Phys J B-Condens Matter Complex Syst 85(1):1–22. Springer

Hlinka J, Hartman D, Jajcay N, Vejmelka M, Donner R, Marwan N, Kurths J, Paluš M (2014) Regional and inter-regional effects in evolving climate networks. Nonlinear Process Geophys 21(2):451–462. Copernicus GmbH

Malik N, Bookhagen B, Marwan N, Kurths J (2012) Analysis of spatial and temporal extreme monsoonal rainfall over South Asia using complex networks. Clim Dyn 39(3–4):971–987. Springer

MacQueen J et al (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Berkeley, vol 1, no 14, pp 281–297

Marwan N, Romano MC, Thiel M, Kurths J (2007) Recurrence plots for the analysis of complex systems. Phys Rep 438(5–6):237–329. doi = 10.1016/j. physrep.2006.11.001, ISSN = 03701573

Monahan AH, Fyfe JC, Ambaum MHP, Stephenson DB, North GR (2009) Empirical orthogonal functions: The medium is the message. J Clim 22(24):6501–6514

Reynolds RW, Rayner NA, Smith TM, Stokes DC, Wang W (2002) An improved in situ and satellite SST analysis for climate. J Clim 15(13):1609–1625

Rheinwalt A, Marwan N, Kurths J, Werner P, Gerstengarbe F-W (2012) Boundary effects in network measures of spatially embedded networks. EPL (Europhys Lett) 100(2):28002. IOP Publishing

Romano MC, Thiel M, Kurths J, Mergenthaler K, Engbert R (2009) Hypothesis test for synchronization: twin surrogates revisited. Chaos (Woodbury, N.Y.) 19(1):015108. doi = 10.1063/1.3072784

Thiel M, Romano MC, Kurths J, Rolfs M, Kliegl R (2006) Twin surrogates to test for complex synchronisation. Europhys Lett (EPL) 75(4):535–541. doi = 10.1209/epl/i2006-10147-0, ISSN = 0295-5075

Thiel M, Romano MC, Kurths J, Rolfs M, Kliegl R (2008) Generating surrogates from recurrences. Philos Trans Ser A Math Phys Eng Sci 366(1865):545–557. doi = 10.1098/rsta.2007.2109, ISSN = 1364-503X

Tsonis AA, Roebber PJ (2004) The architecture of the climate network. Phys A Stat Mech Appl 333:497–504. Elsevier

Tsonis AA, Swanson KL, Roebber PJ (2006) What do networks have to do with climate? Bull Am Meteorol Soc 87(5):585–595

Trenberth KE, Stepaniak DP, Caron JM (2000) The global monsoon as seen through the divergent atmospheric circulation. J Clim 13(22):3969–3993

Webster PJ, Magana VO, Palmer TN, Shukla J, Tomas RA, Yanai M, Yasunari T (1998) Monsoons: processes, predictability, and the prospects for prediction. J Geophys Res Oceans (1978–2012) 103(C7):1445114510. Wiley Online Library

Webster PJ, Moore AM, Loschnigg JP, Leben R (1999) Coupled ocean-atmosphere dynamics in the Indian Ocean during 1997–98. Nature 401(6751):356–360. Nature Publishing Group

Wiedermann M, Donges JF, Heitzig J Kurths J (2013) Node-weighted interacting network measures improve the representation of real-world complex systems. EPL (Europhys Lett) 102(2):28007. IOP Publishing

Yamasaki K, Gozolchiani A, Havlin S (2008) Climate networks around the globe are significantly affected by El Nino. Phys Rev Lett 100(22):228501. APS

Yatagai A, Kamiguchi K, Arakawa O, Hamada, A, Yasutomi N, Kitoh A (2012) APHRODITE: constructing a long-term daily gridded precipitation dataset for Asia based on a dense network of rain gauges. Bull Am Meteorolog Soc 93(9):1401–1415. American Meteorological Society