Comparison of Spatio-Temporal Prediction Approaches of Point-Referenced Environmental Data

Nina Dorffer [1], Julian Bruns [2], Andreas Abecker [3], Stefan Lossow [4]

Abstract: Due to climate change and the effect on human health, there is an urgency to observe and understand the environment. To achieve this goal, knowledge about the development of environmental parameters over time and space is necessary. The analysis of the underlying data can therefore be done with spatial, temporal or even spatio-temporal methods. These methods can also be combined: First spatial,then temporal analyses and vice versa. In this work, we compare the effect of the decision whether to first analyze the data spatially, temporally or both simultaneously. We chose temperature data in Baden-Württemberg, yearly and monthly aggregated, for our comparison.

Keywords: Environmental Computer Science; Spatio-Temporal Statistics; Evaluation

1 Introduction

The environment is everything around us. It is controlled by physical, chemical and other natural forces. Since humans embraced industrialisation different environmental disasters occurred. Two of the most common risks to human health in the developed world are introduced by air pollution and heat stress. Air pollution is defined as an emission of substances in form of aerosols or gas like sulphur dioxide or nitrogen oxides [EA20]. The effect of heat stress is investigated i.e. by [Mu19] in the central and southern parts of Germany. The environment is an important part of human living. Changes in any environmental parameter can cause consequences on human health. Consequently, there is a need to observe our environment.

Environmental data sets often consist of information about the date and location of observation. Based on these information, the analysis of this data can be spatial, temporal or even spatio-temporal. “Spatio-temporal aspects are crucial in order to manage and mine data, to index and retrieve information, and finally to discover and visualize knowledge” [IRT18]. In the last decade, the temporal and spatial resolution of environmental observations increased [GPH16a]. This led to a bigger amount of spatio-temporal datasets. As a result, the issue of spatial, temporal and spatio-temporal analyses of different environmental parameters seems necessary.

[1] Disy Informationssysteme GmbH, Ludwig-Erhard-Allee 6, 76131 Karlsruhe nina.dorffer@disy.net
[2] Disy Informationssysteme GmbH, Ludwig-Erhard-Allee 6, 76131 Karlsruhe julian.bruns@disy.net
[3] Disy Informationssysteme GmbH, Ludwig-Erhard-Allee 6, 76131 Karlsruhe andreas.abecker@disy.net
[4] Disy Informationssysteme GmbH, Ludwig-Erhard-Allee 6, 76131 Karlsruhe stefan.lossow@disy.net

doi:10.18420/inf2022_129

One restriction of most environmental ground-measured data is that they are mostly measured at fixed points, not over a larger area. In fact, an often appearing request by for example city planners is the availability of data over a whole study area. One way to get such data is from satellites. Usually, the temporal resolution from satellites is currently lower than from ground-stations. A high temporal resolution is needed to calculate a more accurate forecast [SH10]. There are only few possibilities to measure environmental data continously over space and time.

By combining temporal data with a spatial domain it is implied that both variabilities, spatial and temporal, are modelled. As a result, a spatio-temporal analysis is more complicated than just a spatial or temporal analysis. Traditionally the dimensions space and time are treated separately [Li14]. In practice, the location attribute or time attribute is often ignored at first. Consequently, the fields of spatial analysis and time series analysis have developed independently much from on another and the spatio-temporal context is often missing. This leads to the effect that the analysis of space-time data is often done conditionally. To analyse space-time data conditionally means, that first the spatial aspects are analysed then temporal aspects or vice versa. In addition to that, there is often a lack of good software to handle, import, export, display and analyse space-time data which might be the reason for this handling of space-time data. However, new software such as the R package space time and text books are starting to fill this gap ([Bi08], [CW15],[WZMC19], [GPH16a] and [GG10]).

All of this leads to a lack of knowledge regarding the practical application of spatio-temporal analyses - conditional or simultaneously. We aim to compare the impact of the difference in practical applications. By using real world, open datasets from the environmental domain, we select suitable methods and compare their results. We focus on well-known and already implemented methods in the space time and similar R packages, as these would be used by most environmental scientists and practioneers.

2 Related Work

There exists an impressive body of work which compares spatial or temporal interpolations as well as predictions. We have chosen to present a small look into this field.

[Go00] interpolated monthly and yearly aggregated rainfall. Three multivariate approaches using elevation and three univariate methods were compared. The univariate method ordinary kriging got the best result only in the case when the correlation between rainfall and elevation was less than 0.75. [LTP08] compared spatial interpolation methods. They interpolated continuous wind speed surfaces in the study area of England and Wales. The best result was achieved by using co-kriging including the elevation data. Key result is that range, variance and input data values are key parameter, when choosing a method. [Ne14] tested an artificial neural network (ANN) for spatial interpolation. Afterwards, the result was compared with common interpolation methods like inverse distance weighting (IDW) and ordinary kriging. The result was worse and the computation time is higher than with common interpolation methods.

There are a lot of studies about the comparison of different methods to forecast a time series. For example, [UCK08] compared three ANN methods. The temperature of two research areas in Turkey were investigated. Each ANN produced usable results. Another example is the study by [St12]. They predicted the daily water temperature of the Moisie River, a watercourse in eastern Canada. They concluded that ANN models are promising to forecast water temperature if it was a long time series. The study by [CD14] compared different ARIMA-models to identify a trend of pollutants, temperature and humidity in India. [Te19] compared the forecast by ANN and ARIMA models by calculating a daily forecast for the output energy of wind turbine. In their paper, ARIMA and SARIMA achieved better results than an ANN. Further studies such as [DH06], the M5 competition [MSA22] or textbooks such as [CM09] provide a starting point ofr a broader view into this topic beyond the environmental sciences.

Most studies dealing with spatio-temporal analysis were presenting a solution for a specific problem. They used different approaches, like hierarchical models, ANN or spatio-temporal interpolation. [Mc05] and [SYH09] used a Bayesian hierarchical model to describe and forecast ozone. [An01] used ANN because there was a lack of data in the past. [PI13] calculated a forecast of ozone concentration in the city centre of Athens with ANN. [TdRM20] present an approach for forecasting a large study area. The time series at each station was decomposed. Trend and seasonality were extrapolated. That approach yields a good result for large areas and long-term forecast. [JK17] compared purely spatial methods with isotropic spatio-temporal and anisotropic spatio-temporal approaches. In the example of this study of irradiance data, the anisotropic approach achieved the best results. [Am20] present ä framework for spatio-temporal prediction of climate and environmental data using deep learning"which is evaluated on simulated as well as real-world temperature data.

Most of the studies presented solutions for a specific problem. The comparison of different methods to interpolate or forecast using space and time is not a subject of broader study to the knowledge of the authors. This paper aims to contribute to fill a gap by providing a starting point in comparing and presenting different methods to forecast a study area. We aim to provide first insights when to use which approach, especially for environmental data.

3 Study Area

For our comparison of the different approaches, we use public data for the federal state of Baden-Würrtemberg. Baden-Württemberg is the third largest federal state of Germany by size as well as by population. This state has a variety of different landscapes like the Black Forest, the Rhine plain, the lake of Constance and the Alps in the south (Baden-Württemberg, n.d.). The Swabian Alb in the east determines the landscape.

Data sets for the study area are provided by LUBW (State Office for the Environment, Measurements and Nature Conservation of the Federal State of Baden-Württemberg) as well as by DWD (German national meteorological service). The data used are open and freely available.

The data sets by DWD are combined with the data sets by LUBW to perform a better analysis. The requirements for the data are, that a raster for validation exists and that there are as many stations as possible. Table 1 shows the number of stations that measure meteorological data in Baden-Württemberg divided by LUBW, DWD and both. In addition to that, it is shown whether a raster by DWD from their open data portal, that can be used to validate, is available or not. The temporal resolution for raster providing precipitation is hourly, daily, monthly, yearly, and for years. The temperature raster is available monthly, yearly and for years. For this work, we assume the raster data from the DWD to be correct and not interpolated. While this may introduce an additional bias into the evaluation, we believe that for an early investigation this is acceptable.

	Precipitation	Temperature	Wind velocity	Wind direction	Air pressure
LUBW stations	26	25	26	26	10
DWD stations	131	100	93	93	43
Number of stations	157	125	119	119	53
Raster available	Yes	Yes	No	No	No

Tab. 1: Available meteorological data from LUBW and DWD

Based on the existing data, we use temperature data for our study from the time period 2009 - 2019. We use monthly as well as yearly aggregated data based on the availability. While precipitation data is also feasible, temperature has a field of investigation has a stronger presence in the literature and the results are therefore easier to compare to the literature.

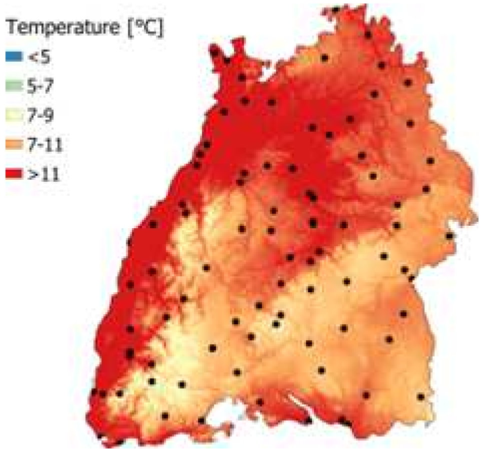

Fig. 1: Temperature raster for the year 2019 and observation stations

4 Methodology and Evaluation

Three different methods are compared in this study. (1) Forecasting using spatial interpolation, then temporal forecasting (S-T), (2) Forecasting using space-time inverse distance weighting (stIDW) and (3) Forecasting using temporal prediction, then spatial interpolation (T-S). For both S-T as well as T-S we compared different approaches for both the spatial and temporal analysis. The best approaches were then chosen and combined in this study. For the temporal models we use an auto-fitted SARIMA model, for the temporal data, we used the auto.kriging function of gstat [GPH16b].

4.1 Forecasting using spatial interpolation, then temporal forecasting (S-T)

In the first step, all previous time steps between 2009 and 2019 are interpolated either yearly or monthly aggregated. The decision, which interpolation method is used, is done by calculating the mean absolute error of the cross-validation. The method resulting in the lowest MAE is chosen. Afterwards, all interpolated raster are laid on top of each other. For each pixel, a time series is created, then for each pixel the value(s) for 2019 is predicted. The workflow can be seen in figure 2. The accuracy of the validation data is calculated for each method. The method with the smallest MAE is chosen to calculate the forecast. To validate the result two different methods are used.

a) In the first validation, the accuracy of the spatial and temporal part is observed individually. We use cross-validation for the spatial anaylsis. The validation of the temporal part is done with a split into training, validation and test data. Therefore the model is trained using data from 2009 – 2016. The validation data used is from 2017 – 2018 and 2019 as evaluation data. This allows us to incorporate a full year with all seasons while still incorporate a large amount of data for training.

b) The result is also validated by calculating the difference to the validation data by DWD. For monthly aggregated data this validation is done for all month. The results are presented for each month and combined by presenting the average.

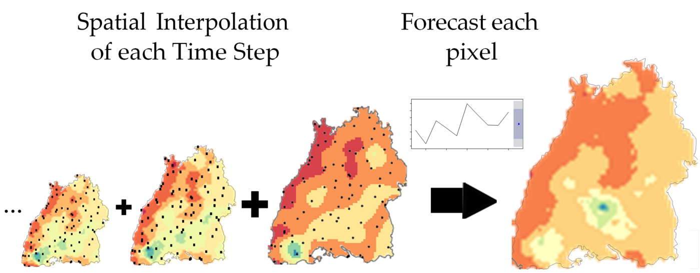

Fig. 2: Workflow to forecast a study area using spatial interpolation, then temporal forecasting

4.2 Forecasting using space-time inverse distance weighting (stIDW)

The method space-time IDW from the package geosptdb is used to calculate the forecast. Space-time IDW is used because the time-intervals were not suitable for space-time kriging in our experiments - during the implementation time, only interpolaton intervalls of less than a day were possible in the package. Based on a preliminary study, the number of neighbours is set to four, the inverse distance power to two and C to one. Therefore the data of the years 2009 till 2018 are used to train the model and to compute a cross-validation. The year 2019 and each month of 2019 was interpolated. To validate the result two different methods are used: a) A leave-one-out cross-validation is calculated and presented. Advantage of this method is, that it is calculated without any information of the year 2019. b) The result is also validated by calculating the difference to the validation raster by DWD.

4.3 Forecasting using temporal prediction, then spatial interpolation (T-S)

The combination is done the other way round in comparison to the previously presented approach for S-T. For each station of the study area values of the wanted time are predicted. The forecast method is chosen by calculating the accuracy of the model for the validation data. The model with the lowest MAE is chosen. Afterwards, the result(s) are interpolated, see figure 3. To decide which method is used to interpolate the MAE of the cross-validation is used. The validation was calculated the same way as for the previous presented method (S-T).

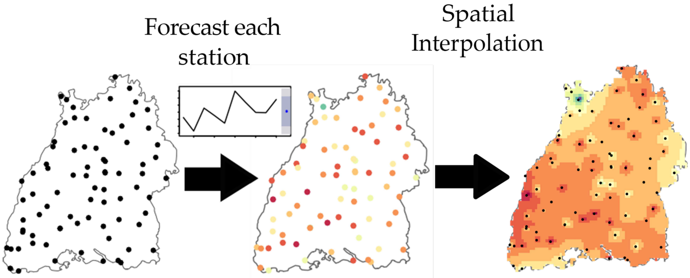

Fig. 3: Workflow to forecast using temporal prediction, then spatial interpolation

5 Results

Firstly, the results for the forecast of the year 2019 are visualized. The test validation leads to the result, that T-S has a low maximal error fig. 4 a. On the other side, the raster validation has a different result. S-T has the lowest MAE and space-time IDW has the lowest maximal error fig. 4 b.

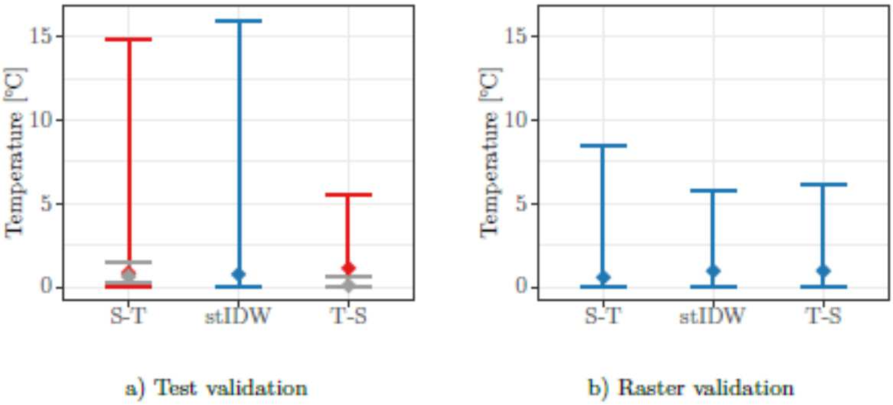

Fig. 4: Comparison of method to forecast of a study area for yearly aggregated temperature. In grey is the temporal error, in red the spatial error and in blue the error of the combined methods.

In fig. 5 the result of the test validation and raster validation of monthly aggregated temperature is presented. In fig. 5a the accuracy of the methods does not differ a lot from each other. It can be said, that the result of the raster validation (fig. 5b) differs from the result of the test validation. In the raster validation, S-T has the lowest maximal error and T-S the lowest mean absolute error. The method space-time IDW has the worst MAE.

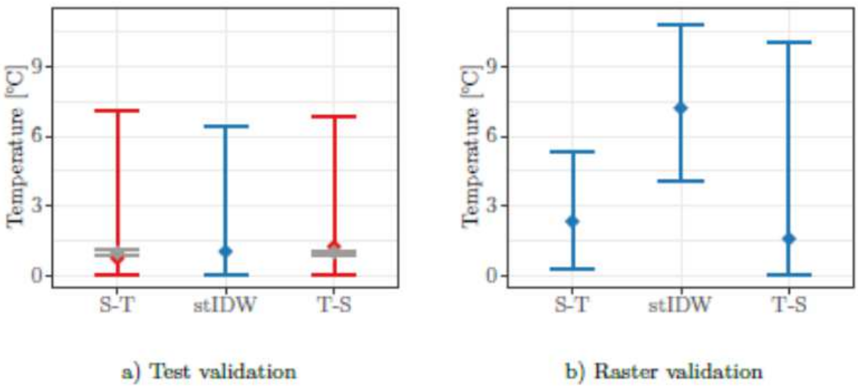

Fig. 5: Comparison of method to forecast of a study area for monthly aggregated temperature. In grey is the temporal error, in red the spatial error and in blue the error of the combined methods.

The results for the raster validation for each month are shown in fig. 6. It can be seen, that the space-time IDW has a MAE about 15 °C in the summer months June, July and August. Reason for that, it that the space-time IDW does not model seasonality. The time part is modelled with an autoregressive model. The methods S-T and T-S have over the year a MAE lower than 5 °C. However it is noticeable, that the maximal error of T-S can get very high, like in March and April. On the other hand, S-T has the largest maximal error in May with about 6 °C (fig. 6).

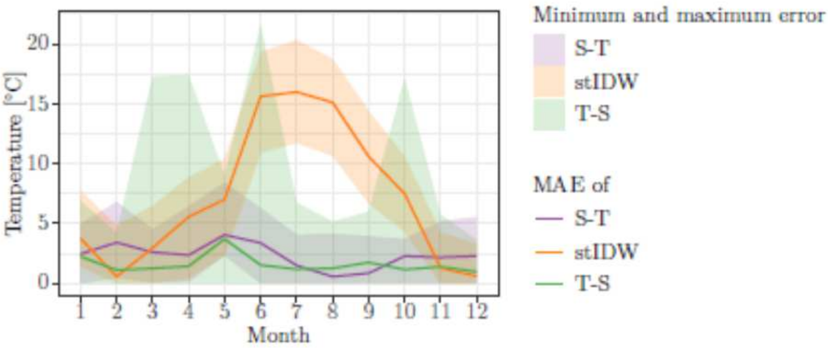

Fig. 6: Raster validation methods to forecast temperature in a study area for monthly aggregated temperature

6 Discussion

Table 2 contains a summary of the results. It is shown which method achieved the lowest MAE, the lowest minimum error and the lowest maximum error. This information is given for the result of the test validation and raster validation. A dash symbolizes that all methods have almost the same result. It can be seen, that results of the test validation and raster validation are not similar. In the test validation the result is often uncertain because all methods led to similar accuracies. Considering the raster validation S-T achieved in most cases the best result. That means the test validation does not help to find the best method to forecast the study area. In fig. 6 it can be seen, that in these cases the reduction methods, S-T and T-S, remains constant and achieves good results in comparison to the extension method space-time IDW.

	Test Validation			Raster Validation		
Data Set	MAE	Max Error	Min error	MAE	Max Error	Min error
Yearly temp	-	T-S	-	S-T	stIDW	-
Monthly temp	S-T	stIDW	-	T-S	S-T	T-S

Tab. 2: Methods with best accuracy in cross-validation and validation to forecast a study area

7 Conclusion

"When data are sparse in space but dense in time, [...] the analysis can be done within the framework of multivariate time series. [...] When data are dense in space and sparse in time, [...] one can work in a multivariate geostatistical setting [...]. In the case, that data are dense in time and space, space-time methods should be used." [Ne17]. While this rule is broadly helpful, it also does not specifiy what the appropiate denseness - the amount of measurements - has to be. In this study we aimed to provide more insight when to use which approach. Three approaches to forecasting a study area were analysed. To interpolate first and make a forecast on each pixel (S-T) led to good forecast results in most investigated cases. It appeared that this approach needed conspicuously long execution time. In comparison, the S-T method needed about 30 times more time than the T-S. And the T-S needed about 30 times more time than space-time IDW based on the work conducted in this paper. Consequently, at this stage, it is either possible to get a result, that is not accurate or to wait until the calculation is done. However, given our dataset, we can say for studies focused on measurement stations - providing dense measurements in time, but not in space - we would advise to first use spatial analysis, followed by the temporal analysis. However, a more in-depth investigation is needed to provide broadly transferable insights.

In this work we used data with a high temporal resolution and a relatively small spatial resolution. This is typical for sensor stations and can also be prevalent in most IoT (Internet of Things) scenarios. We expect this to be a growing use case for spatio-temporal analyses as the IoT use case can be seen in most smart city scenarios - i.e. in the water sector. As sensors are getting cheaper , we will see more application scenarios for those. And while this will lead to an increased spatial coverage, the measurement times will also be reduced to hourly or even measurements every second. But it can also be said, that the method space-time IDW is an improper method for especially monthly aggregated environmental data. That is because, for example, the seasonality is not modelled. The time is modelled using an autoregressive model, which is not sufficient for this use case and therefore most studies in the environmental fields in the next few years.

Another result would have been likely if we investigated the opposite resolutions of the data - low temporal resolution and high spatial resolution. This can be most often seen in the remote sensing field. A follow up investigation to see if the S-T, T-S or spatio-temporal would lead to better results would be of high interest to the authors. Especially of interest is to show or even proof what the conditions are for the better performance of each approach.

For future workother spatio-temporal interpolation methods should be compared, like space-time kriging. This method was not considered in this study because it requires the temporal aggregation not to exceed daily values. In our opinion, not only should more existing methods be compared, but there is also a need for the development of new approaches of spatio-temporal analysis. There do exist a manifold of research paper in regard to natural language processing, even more in regard to image detection (to the benefit of remote sensing), but the use of artifical intelligence for environmental problems with spatio-temporal data there is still a research gap. Most methods are still provided by statistics - both old and new - but a broader toolset can only benefit the field.

8 Acknowledgement

We would like to thank the reviewers for their intensive as well as constructive feedback. This work was funded by the BMU within the project Nitrat-Monitoring4.0 - Intelligente Systeme zur nachhaltigen Reduzierung von Nitrat im Grundwasser (NiMo 4.0), FKZ 67KI2048.

Bibliography

[Am20] Amato, Federico; Guignard, Fabian; Robert, Sylvain; Kanevski, Mikhail: A novel framework for spatio-temporal prediction of environmental data using deep learning. Scientific reports, 10(1):1–11, 2020.

[An01] Antonić, Oleg; Križan, Josip; Marki, Antun; Bukovec, Dragan: Spatio-temporal interpolation of climatic variables over large region of complex terrain using neural networks. Ecological Modelling, 138(1-3):255–263, 2001.

[Bi08] Bivand, Roger S; Pebesma, Edzer J; Gómez-Rubio, Virgilio; Pebesma, Edzer Jan: Applied spatial data analysis with R, volume 747248717. Springer, 2008.

[CD14] Chaudhuri, Sutapa; Dutta, Debashree: Mann–Kendall trend of pollutants, temperature and humidity over an urban station of India with forecast verification using different ARIMA models. Environmental monitoring and assessment, 186(8):4719–4742, 2014.

[CM09] Cowpertwait, Paul SP; Metcalfe, Andrew V: Introductory time series with R. Springer Science & Business Media, 2009.

[CW15] Cressie, Noel; Wikle, Christopher K: Statistics for spatio-temporal data. John Wiley & Sons, 2015.

[DH06] De Gooijer, Jan G.; Hyndman, Rob J.: 25 years of time series forecasting. International Journal of Forecasting, 22(3):443–473, 2006. Twenty five years of forecasting.

[EA20] Emetere, Moses Eterigho; Akinlabi, Esther Titilayo: Introduction to environmental data analysis and modeling, volume 58. Springer Nature, 2020.

[GG10] Gaetan, Carlo; Guyon, Xavier: Spatial statistics and modeling, volume 90. Springer, 2010.

[Go00] Goovaerts, Pierre: Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. Journal of hydrology, 228(1-2):113–129, 2000.

[GPH16a] Gräler, Benedikt; Pebesma, Edzer J; Heuvelink, Gerard BM: Spatio-temporal interpolation using gstat. R J., 8(1):204, 2016.

[GPH16b] Gräler, Benedikt; Pebesma, Edzer; Heuvelink, Gerard: Spatio-Temporal Interpolation using gstat. The R Journal, 8:204–218, 2016.

[IRT18] Inkpen, Diana; Roche, Mathieu; Teisseire, Maguelonne: , Guest editorial: special issue on Environmental and Geospatial Data Analytics, 2018.

[JK17] Jamaly, Mohammad; Kleissl, Jan: Spatiotemporal interpolation and forecast of irradiance data using Kriging. Solar Energy, 158:407–423, 2017.

[Li14] Li, Lixin; Losser, Travis; Yorke, Charles; Piltner, Reinhard: Fast inverse distance weighting-based spatiotemporal interpolation: a web-based application of interpolating daily fine particulate matter PM2. 5 in the contiguous US using parallel programming and kd tree. International journal of environmental research and public health, 11(9):9101–9141, 2014.

[LTP08] Luo, W; Taylor, MC; Parker, SR: A comparison of spatial interpolation methods to estimate continuous wind speed surfaces using irregularly distributed data from England and Wales. International Journal of Climatology: A Journal of the Royal Meteorological Society, 28(7):947–959, 2008.

[Mc05] McMillan, Nancy; Bortnick, Steven M; Irwin, Mark E; Berliner, L Mark: A hierarchical Bayesian model to estimate and forecast ozone through space and time. Atmospheric Environment, 39(8):1373–1382, 2005.

[MSA22] Makridakis, Spyros; Spiliotis, Evangelos; Assimakopoulos, Vassilios: M5 accuracy competition: Results, findings, and conclusions. International Journal of Forecasting, 2022.

[Mu19] Muthers, Stefan; Niemann, Hildegard; Buchholz, Udo; Grabenhenrich, Linus; Matzarakis, Andreas et al.: Schätzung hitzebedingter Todesfälle in Deutschland zwischen 2001 und 2015. Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz, 62(5):571–579, 2019.

[Ne14] Nevtipilova, Veronika; Pastwa, Justyna; Boori, Mukesh Singh; Vozenilek, Vit: Testing artificial neural network (ANN) for spatial interpolation. Journal of Geology & Geophysics, 3(2):01–09, 2014.

[Ne17] Network, RESSTE et al.: Analyzing spatio-temporal data with R: Everything you always wanted to know–but were afraid to ask. Journal de la Société Française de Statistique, 158(3):124–158, 2017.

[PI13] Papaleonidas, Antonios; Iliadis, Lazaros: Neurocomputing techniques to dynamically forecast spatiotemporal air pollution data. Evolving Systems, 4(4):221–233, 2013.

[SH10] Sallis, Philip John; Hernandez, Sergio: Ensemble interpolation methods for spatio-temporal data modelling. In: 2010 Fourth UKSim European Symposium on Computer Modeling and Simulation. IEEE, pp. 132–135, 2010.

[St12] St-Hilaire, André; Ouarda, Taha BMJ; Bargaoui, Zoubeida; Daigle, Anik; Bilodeau, Laurent: Daily river water temperature forecast model with ak-nearest neighbour approach. Hydrological Processes, 26(9):1302–1310, 2012.

[SYH09] Sahu, Sujit K; Yip, Stan; Holland, David M: Improved space–time forecasting of next day ozone concentrations in the eastern US. Atmospheric Environment, 43(3):494–501, 2009.

[TdRM20] Takafuji, Eduardo Henrique de Moraes; da Rocha, Marcelo Monteiro; Manzione, Rodrigo Lilla: Spatiotemporal forecast with local temporal drift applied to weather patterns in Patagonia. SN Applied Sciences, 2(6):1–19, 2020.

[Te19] Tena García, Jorge Luis; Cadenas Calderón, Erasmo; González Ávalos, Gilberto; Rangel Heras, Eduardo; Mbikayi Tshikala, Alain: Forecast of daily output energy of wind turbine using sARIMA and nonlinear autoregressive models. Advances in Mechanical Engineering, 11(2):1687814018813464, 2019.

[UCK08] Ustaoglu, B; Cigizoglu, HK; Karaca, M: Forecast of daily mean, maximum and minimum temperature time series by three artificial neural network methods. Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling, 15(4):431–445, 2008.

[WZMC19] Wikle, Christopher K; Zammit-Mangion, Andrew; Cressie, Noel: Spatio-temporal Statistics with R. Chapman and Hall/CRC, 2019.