ESTIMATING RAINFALL FOR INDEX-BASED AGRICULTURAL INSURANCE

Arif Albayrak
William Teng
ADNET Systems, Inc.
NASA Goddard Earth Sciences Data and Information Services Center
Code 610.2
Greenbelt, MD 20771

ABSTRACT

Index-based agricultural insurance is a promising alternative for individual farmers who cannot afford traditional insurance based on field inspections for assessing losses. Weather-indexed agricultural insurance, with payouts tied to pre-determined triggers of local indices, requires no field visits, thus significantly reducing costs. To be effective, weather indices need a sufficiently dense network of quality-controlled weather stations. However, in some regions of the world, the number of stations is often limited and located in the main cities and not near farms that are to be insured.

Our study focused on estimating rainfall for areas without rain gauges. For the first phase, we assumed there were no available data from the particular farm to be insured. To obtain daily rainfall estimations for that farm, we applied the neural networks method to interpolate/discover rainfall data, using information from neighboring stations. For the second phase, we assumed there were some limited data available from the farm. To test our method, we used 10 years of rainfall data collected from 21 land-based stations in Iowa, U.S. We started with a small number of stations and systematically increased that number and, thus, the information content of the entire system.

Overall, this study showed an improvement in rainfall estimation, when information content from station data increased. While promising for application to weather-indexed agricultural insurance in some areas, the method used in this study could not help extend insurance coverage to areas too far away from existing stations. Thus, the next step would be to incorporate satellite data to increase the density of rainfall and other measurements. The upcoming Global Precipitation Measurement (GPM) and Soil Moisture Active Passive (SMAP) missions (2014 launches) will extend the current data records for both measurements into the future, as well as provide improved quality and resolution.

KEYWORDS: Index-based insurance, neural network, precipitation

INTRODUCTION

Rainfall-Based Agricultural Index Insurance (RBAII) is one type of weather index insurance for agriculture, linked to indices such as rainfall, temperature, and humidity, instead of actual loss (i.e., crop failure) (Barrett, 2007). Index-based agricultural insurance is a promising alternative for individual farmers who cannot afford traditional insurance based on field inspections for assessing losses. Weather-indexed agricultural insurance, with payouts tied to pre-determined triggers of local indices, requires no field visits, thus significantly reducing costs. To design and implement such an insurance contract, it is critical to have an index that accurately reflects losses (Bryla-Tressler, 2011; Dick and Stoppa, 2011). To be effective, weather indices need a sufficiently dense network of quality-controlled weather stations, from which accurate rainfall estimation, within a given uncertainty range, can be obtained. However, in some regions of the world, the number of stations is often limited, and the stations are mostly located in the main cities and not near farms that are to be insured. Even where gauge data do exist, there are other limitations, including short historical time series, missing data, reading errors, and poor representation of growing conditions because of poorly sited gauges (Bryla-Tressler, 2011; Dick and Stoppa, 2011; Dinku et al., 2009).

One approach to addressing the problem of data sparsity for RBAII is to interpolate the rain amount for the area of interest (without weather station) from neighboring weather stations. Two common techniques are used: (1) Deterministic, such as isohyet, Thiessen polygons, Inverse Distance Weighting (IDW) and (2) geostatistical, such as Trend Surface Interpolation (kriging), with known constant mean, unknown constant mean, and non-stationary mean (Ly et al., 2013; Wagner et al., 2012). While these methods have been implemented for different types of situations, interpolation of the data from neighboring stations has been the least successful for precipitation (Xia et al., 1999).

Three main issues need to be considered in the mathematical modeling of the interpolation. **The first** is the time dependency on seasonal and climatological parameters. An example is given by Tokay et al. (2014), in which they showed changing precipitation correlations for different seasons. **The second** issue is the existence of topographical features between stations, such as water bodies and mountains (Goovaerts, 2000). Complex effects of topography on spatial distribution of precipitation are not well known and difficult to incorporate into the precipitation imputation algorithms (Daly et al., 1994; Thornton et al., 1997; Xia et al., 1999). **The third** issue is the distances between stations (Ahrens, 2005). Tokay et al. (2014) showed, for network with separation distances ranging from 1 km to 150 km on the Delmarva Peninsula, that correlation between stations depended not only on distance but also on rain accumulation periods.

These complex relationships among regressors for precipitation interpolation are problematic. A powerful approach to dealing with these complexities is to use a nonparametric method, with which all the statistical properties of the precipitation can be considered. One such method, neural network (NN), has been successfully used for many different earth science applications. Examples include retrieval of land surface reflectance (Liang et al., 2003), retrieval of precipitation from microwave observations (Leslie et al., 2008), and aerosol optical depth bias correction (Albayrak et al., 2013; Das et al., 2008; Ristovski et al., 2012).

The overall objective of this study is to design a mathematical approach to estimate rainfall amount for a specific geographical location, by using the nearest available data sources. For this initial work, we mainly focused on characterizing rainfall data from three weather stations in Iowa, in order to optimize the running of a neural network (as the mathematical estimator). This approach allows one to reduce (i.e., compensate for) the different types of sensor and algorithm retrieval biases, without needing to know, in advance, the sources of systematic errors (Albayrak et al., 2013). The NN used in this study is a supervised learning algorithm, in which training data comprise the input vectors along with their corresponding target vectors. If the desired output consists of one or more continuous variables, then the task is called NN regression, which is the meaning for the term "NN" used in the remainder of this paper. The following sections provide information on the weather station data used in the study, data analysis methods, NN method using a back propagation algorithm, experimental results, and summary.

DATA COLLECTION AND FORMATTING FOR USAGE

We used Global Historical Climatology Network (GHCN)-Daily data from the National Climatic Data Center (NCDC). This is an integrated database of daily climate summaries from land surface stations across the globe. GHCN-Daily now contains records from over 75,000 stations in 180 countries and territories. Both the record length and period of record vary by station. Temporal coverage of the data ranges from less than a year to more than 175 years, as described in http://www.ncdc.noaa.gov/oa/climate/ghcn-daily/.

We used two different META data for the purpose of automatic retrieval: Station Definitions and Data Set Definitions. In the case of station definitions, we collected the following information for each station: NCDC unique identifier, Station Name, Country Name, State Name, Time Zone, Ground Elevation (meters), Latitude, and Longitude. For specific data sets, we obtained the variable definitions from the second META data with five core elements: Precipitation (tenths of mm), Snowfall (mm), Snow Depth (mm), Maximum Temperature (tenths of degree C), and Minimum Temperature (tenths of degree C).

This process can be summarized as retrieving station ID numbers from META data and filtering them according to state. Then, this information is fed into the retrieval program for a given date interval. Data are downloaded using a Python library called Ulmo, which provides a simple and fast access to public hydrology and climatology data.

DATA ANALYSIS

We collected data from 128 stations in Iowa (Fig. 1a), from which we selected 21 with continuous data for 10 years, from 2000 to 2010 (Fig. 2b). From the selected 21 stations, we designated one as the "target" for the study and selected three stations within a 100-km radius from the target (Fig. 2c). These four stations were used for our experiments.

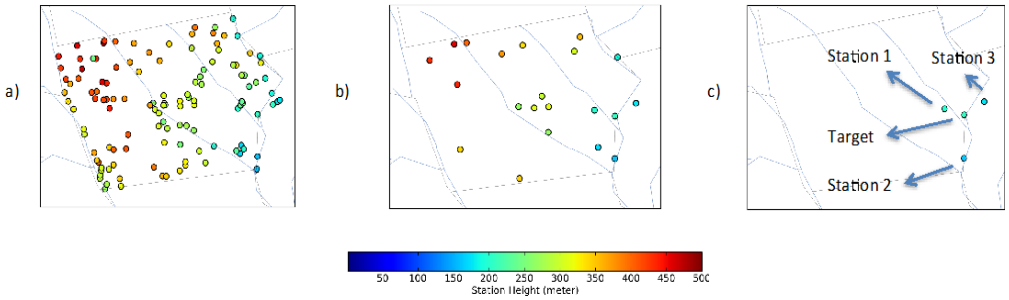

Figure 1. Iowa station locations of Global Historical Climatology Network (GHCN)-Daily data from the National Climatic Data Center (NCDC): (a) All available stations, (b) stations with 10-year continuous data, (c) final four stations within 100 km of each other, used in the study.

In Figure 1c, the target represents the farm location where rain amount is to be estimated. Data from the other three stations are used for interpolation purposes. The distances between the target and stations 1, 2, and 3 are 48,504 m, 90,983 m, and 55,437 m, respectively.

Proper use of NN requires a preliminary analysis of the data sets. Specifically, for the precipitation problem, relations between stations have to be determined. For this purpose, three aspects of the data were analyzed: Cross-Correlation Coefficient (c-cc); Data Distribution; and Precipitation Accumulation.

As a first step, Cross-Correlation Coefficient (c-cc) values between stations were calculated. In addition, cc_c's time lag effects were studied to determine if there was any time latency between precipitation events. Table 1 summarizes changing c-cc's over a 10-year period, between the stations and between stations and the target (farm).

Table 1. Cross correlation coefficient values over a 10-year period between stations and between the target (farm) and stations. S=station T=target

	S1 vs S2		S3 vs S1		S3 vs S2		T vs S1		T vs S2		T vs S3	
	L0	L1	L0	L1	L0	L1	L0	L1	L0	L1	L0	L1
2000	0.57441	0.10582	0.14878	0.54096	0.329638	0.63355	0.1916	0.5743	0.40474	0.61763	0.8082	0.2009
2001	0.57762	0.15497	0.11481	0.69229	0.235935	0.57179	0.23736	0.79819	0.28195	0.5733	0.7342	0.111
2002	0.52028	0.14567	0.30828	0.36592	0.209968	0.43885	0.39989	0.52506	0.4081	0.59228	0.6554	0.2278
2003	0.49094	0.15701	0.24654	0.7269	0.362694	0.5837	0.24897	0.81087	0.21113	0.65099	0.8097	0.2451
2004	0.50545	0.21985	0.18568	0.55974	0.234171	0.61408	0.26625	0.66473	0.20542	0.47476	0.7733	0.2005
2005	0.55526	0.16912	0.31813	0.53767	0.221985	0.39422	0.38404	0.59563	0.35088	0.50849	0.6963	0.1524
2006	0.63181	0.16561	0.17703	0.69004	0.201677	0.60325	0.2775	0.70257	0.33176	0.54503	0.7375	0.1002
2007	0.70608	0.20662	0.27347	0.3428	0.339297	0.34596	0.54713	0.54113	0.63036	0.46879	0.5497	0.1905
2008	0.61986	0.10644	0.26951	0.49661	0.482857	0.62427	0.27258	0.65464	0.45467	0.64041	0.703	0.2592
2009	0.66927	0.20643	0.33391	0.69414	0.311923	0.54483	0.3696	0.78294	0.41392	0.49355	0.7096	0.2923
2010	0.54348	0.18053	0.33061	0.65103	0.311023	0.41862	0.33703	0.70054	0.36272	0.48687	0.8688	0.1629
Overall	0.57611	0.16613	0.25004	0.5466	0.299805	0.51407	0.33075	0.6542	0.37461	0.53622	0.7183	0.205

This analysis allowed us to establish relationships between stations. For example, a comparison between Station 1 (S1) and Station 2 (S2) showed the average or overall c-cc value of 0.57 in lag 0. When S3 and S1 were compared, an overall c-cc value of 0.54 was obtained for lag 1, indicating a 1-day shift between the stations. It should be noted that c-cc's were calculated for each year separately as well. While cc-c's did change from year to year, they were consistent in terms of amplitude and lag effects.

As a second step, data distribution in each station was analyzed, considering scatter plots. Figures 2a and 2b summarize, respectively, the scatters with original data and after shifting stations 2 and 3 by one day.

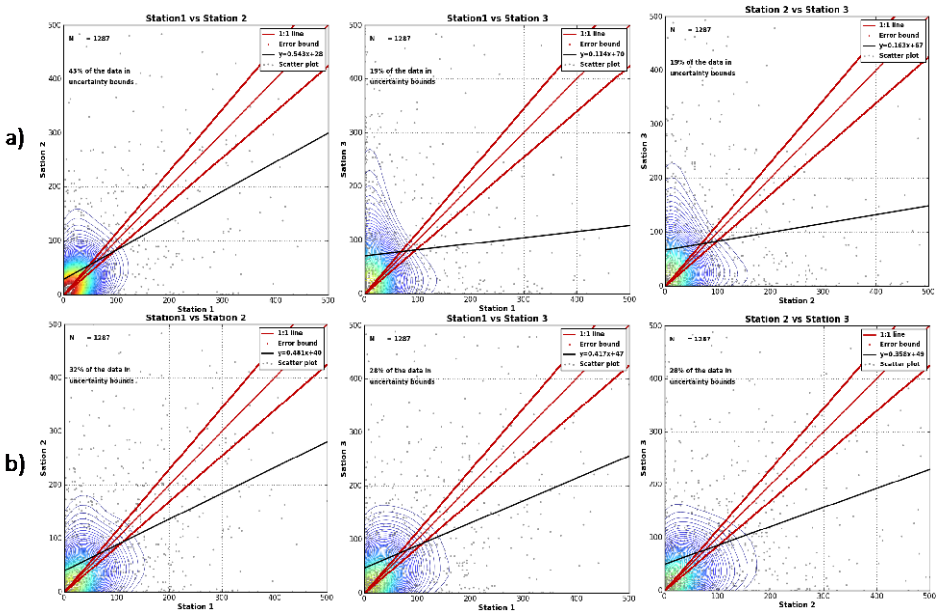

Figure 2. Scatter plots between stations: (a) Before and (b) after lag effect was removed.

In each scatter plot, the cones (the three red lines with the middle one representing the 1-1 line) are defined to detect the change in terms of percent of data that falls into the cone. In summary, shifting of data corrects the distribution as seen in Figure 2 (middle and right columns) and data density decreases between S1 and S2 (left column).

As a third step, rain amounts between stations were analyzed. Because of topographical effects, it is common to have changing amounts of rain accumulation over a time period. One of the best ways to obtain such information is to analyze rolling means (Fig. 3).

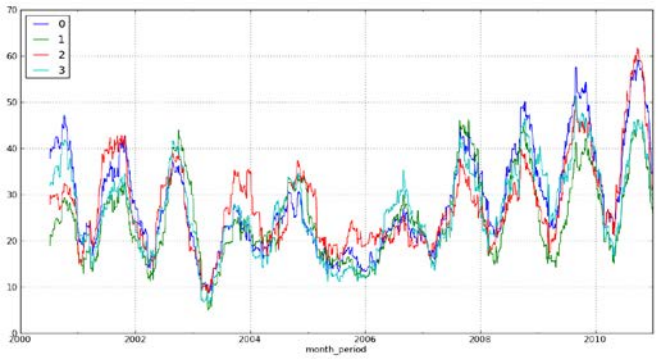

Figure 3. 200-day rolling means for the target and other three stations. (0=target, 1=station 1, etc.)

From Figure 3, one can infer that there are no big differences between the stations. Also, seasonal dependency can be easily seen. However, some type of exception occurred between 2006 and 2007, in which rain amounts decreased considerably.

Based on these preliminary analyses, the following elements needed to be considered before using NN: (1) Lag effects have to be removed, (2) seasonal dependency has to be resolved, and (3) correlations between stations must be reflected at the farm location. These elements are considered in the Experiments and Results section.

DATA SET PREPARATION AND METHODS

We used precipitation measurements from stations as our primary data, and, as auxiliary data, we considered

temperature. In order to use these measurements in NN, they were collected in a k by n matrix, where the first column contained the target values, the next three columns contained the station data, and the final columns contained the temperature measurements:

$$\begin{pmatrix} y_1 & x_{11} \dots & x_{1n} \\ \dots & \dots & \dots \\ y_k & x_{k1} \dots & x_{kn} \end{pmatrix}, \quad (1)$$

In equation 1, k is the number of data points, and n is the dimension of each feature vector. The latter is an n by1 array that encodes measurements of an object in a training data set. Each feature can be described as an independent variable, also known as a regressor. Once the data set is ready, the next step is to apply the NN method (Albayrak et al., 2013). NNs are nonparametric function approximators that are used to learn input–output mappings. Each problem solved with NN requires two main components: (1) A network architecture and (2) training data where known input and corresponding output data vectors are provided. In this study, the NN architecture was based on feed-forward neural networks (FNN) (Albayrak et al., 2013). Each FNN contains four main components: Input vectors $\vec{p}$, transfer function f, weights and biases W, and output vectors $\vec{a}$, as described in Eq. (2):

$$\vec{a} = f(W\vec{p}+\mathrm{b}) \quad (2)$$

One important concept in using NN is the learning rule (LR) or training algorithm, which is a procedure for modifying the weights and biases of a network, in order to move the network outputs closer to the target (Albayrak et al., 2013). In NN, as the inputs are applied to the network, the network outputs are compared to the targets. The Learning Rule is then applied to derive the coefficients that are used to adjust the weights and biases of the network. In other words, in Eq. (2), weight matrix W and input vector $\vec{p}$ are optimized in such a way that the error is minimized between $\vec{a}$ and target.

EXPERIMENTS AND RESULTS

The NN method was evaluated at two experimental stages, based on available information. In the first stage, in order to set a base for comparison, all data, including those from the target station, were used (Fig. 1c). For the NN calculations, we considered different sets of regressors: (a) Original precipitation data; (b) lag-effect-corrected precipitation data; and (c) lag-effect-corrected precipitation data with temperature. Scatter plots were used to evaluate the results between target and estimated rainfall amounts (Fig. 4). We supported these plots by providing additional statistical measures, which are depicted in Figure 4 as 1:1 (red line); linear regression (black line); the expected error (EE) cone (red dotted lines); and data number density distribution (color contours). Results for NN approximation showed that the regressor set (c) yielded the best approximations. We arranged the EE cones in such a way that at least 60% of the lag-effect-corrected data with temperature were located within the cone. Then, keeping the cone shape the same, we calculated the percentages of data points inside the defined cone for the three sets of regressors: 56% for (a), 58% for (b) and 60% for (c). This experiment showed that temperature is an important regressor; as a result, we decided to use the regressor set (c) for the rest of the study.

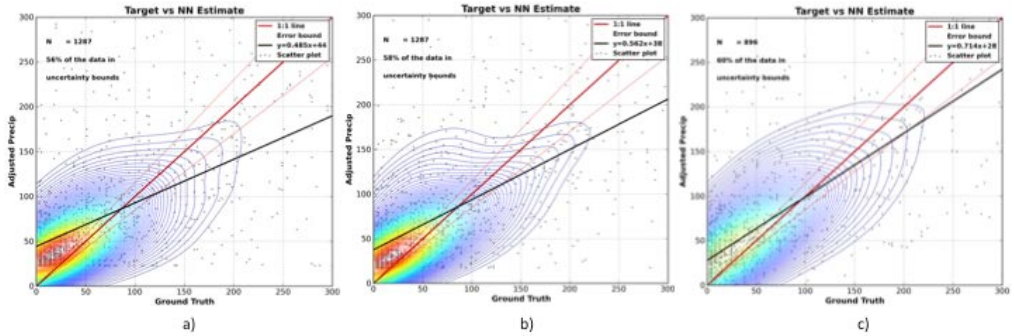

Figure 4. Scatter plots of NN approximations versus the target data for the base case with the three sets of regressors: (a) original precipitation data, (b) lag-effect-corrected precipitation data, (c) lag-effect-corrected precipitation + temperature data.

In the second stage, we prepared two data sets. For the first one, we assumed that no data were available from the target farm and, instead, used weighted-average data from the other three stations (within a 100-km radius) as target data. Weights are adjusted according to distance and correlation coefficients between the stations (Fig. 5).

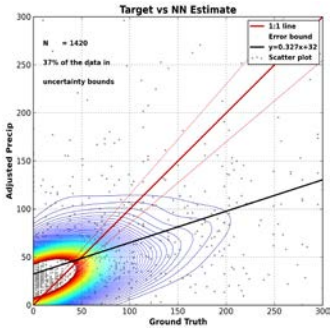

Figure 5. Scatter plot of NN estimates vs. original target data (assume no data were available from target for NN).

For the second case, limited available data from 2000-2001 were used. Compared with the base experiment, where all data, including those from the target station, were considered, the results from the weighted-average estimation were biased, with most of the data outside the EE cone. Only 23% of the observations were inside the cone, compared to 60% for the base, which uses temperature as regressor. The second experiment, which used limited available data, offered improved results, compared to the no-data case, increasing the number of observations within the NN cone from 23% to 46%. These results suggest that actual information from the stations, even incomplete or limited, is preferable to an average from distant stations. Furthermore, additional information (regressors) is required to obtain better approximations. Currently, we are continuing with multi-station experiments.

SUMMARY

In this ongoing study, we designed a mathematical approach using Neural Networks (NN) to estimate rainfall amount for a specific geographical location, by using the nearest available data sources. For this initial phase, we focused on optimizing the running of a NN as the supervised mathematical estimator, by characterizing rainfall data from three weather stations in Iowa. From this analysis, we designed the NN regressors. For the experiments, we considered two different scenarios, based on available information: (1) No available data from the target location and (2) limited available data from the target location. By including limited data, we observed a 23% improvement in estimation results, when running NN with regressors that included additional information such as temperature. That an increase in information content leads to an improvement in estimation results strongly suggests the potential benefits of incorporating satellite data in our methodology for rainfall estimation for index-based agricultural insurance. We plan to make full use of the data from the upcoming Global Precipitation Measurement (GPM) and Soil Moisture Active-Passive (SMAP) missions.

Acknowledgement

We would like to thank Dr. Andrey Savtchenko of the NASA GES DISC for valuable discussions.

REFERENCES

Ahrens B., 2005, Distance in spatial interpolation of daily rain gauge data, *Hydrol. Earth Sys. Sci. Discussions*, 2, 1893–1923.

Albayrak A., J. Wei, M. Petrenko, C. Lynnes, and R.C. Levy, 2013, Global bias adjustment for MODIS aerosol optical thickness using neural network, *J. Appl. Remote Sens.,7(1)*, doi:10.1117/1.JRS.7.073514.

Barrett, C., et al., 2007, Index insurance for climate risk management & poverty reduction: This paper is a policy distillation adapted from IRI Technical Report 07-03 *Working Paper - Poverty Traps and Climate Risk: Limitations and Opportunities of Index-Based Risk Financing*.

Bryla-Tressler D. et al., 2011, Weather index insurance for agriculture: Guidance for development practitioners, *Agriculture and Rural Development Discussion Paper 50*, The International Bank for Reconstruction and Development / The World Bank.

Das, D., V. Radosavljevic, S. Vucetic, Z. Obradovic, 2008, Reducing need for collocated ground and satellite based observations in statistical aerosol optical depth estimation, *IEEE Int'l. Geosci. Remote Sens. Symp.*, Vol. 2, pp. II-879–II-882, doi: 10.1109/IGARSS.2008.4779135

Dick W. and A. Stoppa, 2011, Weather Index-based Insurance in Agricultural Development; A Technical Guide, by the International Fund for Agricultural Development (IFAD), ISBN 9789290722762.

Dinku, T., et al., 2009. Designing index-based weather insurance for farmers in AdiHa, Ethiopia, Report to Oxfam America, Int'l. Res. Inst. for Climate and Society, Columbia Univ., 82 pp.

Goovaerts, P., 2000, Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall, *Journal of Hydrology, 228,* 113–129, http://dx.doi.org/10.1016/S0022-1694(00)00144-X.

Leslie R.V., et al., 2008, Neural network microwave precipitation retrievals and modeling results, *Proc. SPIE, 7154*, 715406, http://dx.doi.org/10.1117/12.804815.

Liang S. et al., 2003, Estimation and validation of land surface broadband albedo and leaf area index from EO-1 ALI data, *IEEE Trans. Geosci. Rem. Sens., 41(6),* 1260–1267, doi:10.1109/TGRS.2003.813203.

Ly, S., C. Charles, and A. Degré, 2013, Different methods for spatial interpolation of rainfall data for operational hydrology and hydrological modeling at watershed scale: A review, *Biotechnol. Agron. Soc. Environ. 17*, 392-406.

Ristovski K., S. Vucetic, and Z. Obradovic, 2012, Uncertainty analysis of neural-network-based aerosol retrieval, *IEEE Trans. Geosci. Rem. Sens., 50*(2), 409–414, http://dx.doi.org/10.1109/TGRS.2011.2166120.

Tokay A, et al., 2014, An experimental study of spatial variability of rainfall, *Journal of Hydrology*, AMT, doi:10.1175/JHM-D-13-031.1.

Wagner P.D., P. Fiener, F. Wilken, S. Kumar, and K. Schneider, 2012, Comparison and evaluation of spatial interpolation schemes for daily rainfall in data scarce regions, *J. Hydrology, 464–465*, 388–400, http://dx.doi.org/10.1016/j.jhydrol.2012.07.026.

Xia Y, P. Fabian, A. Stohl, M. Winterhalter, 1999, Forest climatology: Estimation of missing values for Bavaria, Germany, *Agricultural and Forest Meteorology, 96*, 131–144.