Inferring missing climate data for agricultural planning using Bayesian networks

Leonel Lara-Estrada[1]* & Enrique Sucar[2]

[1] Research Unit Suitability and Global Change. Universität Hamburg, Grindelberg 5, 20144 Hamburg, Germany. Leonel_larae@hotmail.com

[2] Instituto Nacional de Astrofísica, Óptica y Electrónica, Luis Enrique Erro # 1, Tonantzintla, 72840 Puebla, Mexico.

June 2016

Abstract. Climate data availability has a key role in development processes of policies, services, and planning in the agriculture sector. However, a common problem is the lack of data at the spatial or temporal resolution required, or missing variables. In this work we propose a Bayesian Network approach to generate data for those variables that present incomplete data series, maintaining the consistency among variables. We first determine which of the measured variables are the best ones to use as proxies for the unmeasured one based on correlation analysis and availability. Then, based on these variables, we built a Bayesian network model to estimate the variable of interest. Based on this model, the most probable values of the missing variable are estimated via probabilistic inference. We used as a case study the relative humidity (RH) of the driest month, which is one of main variable-indicators to describe the suitability of the land for coffee production. A BN model was built to estimate RH and evaluated for all Central America and Southern Mexico, showing a very good performance over several metrics.

1 Introduction

Climate data availability has a key role in development processes of policies, services, and planning in the agriculture sector. However, a common problem is the lack of data at the spatial or temporal resolution required, or gaps in the data series, or missing variables, or all them combined [15,19,20]. The access to climate information is determinant for agricultural planning. For example, in the case of Central America, crop modeling using process-based models commonly requires climate data in daily time steps, which are hardly available for all the region [18,20]. Even if the data would be available among different climate data sources, possible difference in resolution and origin produce inconsistency issues among the datasets making their combination unsuitable [13]. Therefore, several solutions for reconstructing missing data series using different approaches have been proposed. These include the regularized EM algorithm for Gaussian data [15], empirical orthogonal functions [16], the group method of data handling [1], and others. Frequently many of these approaches have limitations to deal with data uncertainty (from measurements, processes, etc.) and the sort of possible missing information in real time series records and even modeled data. Bayesian networks (BNs) can handle these kind of conditions. However, few attempts have been done to show the advantages of BNs in climate science. Previous work includes the application of BNs for weather forecasting [3] and the exploration of the dependencies among climate variables [4].

We propose a Bayesian Network approach to generate data for those variables that present incomplete data series, maintaining the consistency among variables. We first determine which of the measured variables are the best ones to use as proxies for the unmeasured one based on correlation analysis, and also on availability. Then, based on these variables, we built a Bayesian network model to estimate the variable of interest. Finally, we estimate the most probable values via probabilistic inference.

We used as a case study the relative humidity (RH) of the driest month, which is one of the main variable-indicators to describe the suitability of the land for coffee production. A BN model was built to estimate RH and evaluated for all Central America and Southern Mexico, showing a very good

performance over several metrics. The results confirmed that the estimated RH maintains the consistency with the other variables as in the training data. This means that the estimated RH keeps the same relationship with the other variables, which is one of the major concerns in modeling climate data.
The main contribution of this work is a novel application of Bayesian networks in estimating climate data for agricultural planning, an important and unexplored domain for the application of probabilistic graphical models.

The rest of the document is organized as follows. First, we review previous work on estimating missing data in BNs in general, and in agriculture in particular. Then we described the proposed methodology, including variable selection and model construction. Next, we present the experiments and results, as well as the potential applications of the model. We finalize with a summary and directions for future work.

2 Related Work

2.1 Handling missing data in BNs

When learning a BN from data, a common situation is to have incomplete data. There are two basic cases [17]:

Missing values: there are some missing values for one or more variables.
Hidden nodes: a variable or set of variables for which there is no data at all.

For dealing with missing values, there are several alternatives:

1. Eliminate the registers with missing values.
2. Consider a special unknown value.
3. Substitute the missing value by the mode of the variable.
4. Estimate the missing values based on the values of the other variables in the corresponding register, based on probabilistic inference.

For hidden nodes, the most common approach to estimating their parameters is based on the *Expectation–Maximization* (EM) technique [17].

A technique for error detection and missing data estimation based on Bayesian networks was proposed in [8]. The algorithm starts by building a model of the

dependencies between sources of information (variables) represented as a Bayesian network. Subsequently, the validation is done in two phases. In the first step, potential faults are detected by comparing the actual value with the one predicted from the related variables via propagation in the Bayesian network. In the second phase, the real faults are isolated by constructing an additional Bayesian network based on the Markov blanket property. Using the BN model, missing values can be estimated based on the values of the variables in the Markov blanket of the missing one. In contrast with this work, in our approach, we do not build a BN model for all the variables in the problem but instead look for those variables that can be used as proxies to estimate the missing variable, and based on these variables we build the BN model.

2.2 Estimating missing data in agriculture

Missing data is a major problem in the agricultural sector, in applications such as the definition of policies or implementation of programs at national or regional scales, or farming planning at local or farm levels. Using incomplete information during such processes leads to a misrepresentation of the phenomena under analysis, and completing the missing data has economic, technical, and timing costs [9,12]. Some standard procedures have been used in the sector. For example, the National Agricultural Statistics Service (NASS) that implements the Agricultural Resource Management Survey in the USA, used conditional averages after the elimination of outliers, and if the information is insufficient national averages were used. However, the NASS is in the process of improving the procedures to deal with missing information [12]. In this sense, academia has proposed several methods, such as combining surveys and satellite information [6], spatial interpolations [16], using proxy variables [10], reducing the resolution of analysis, and others. These sort of possibilities are promising; however, in practice institutions and individuals have constraints of different nature to implement these methods. In developing countries, access to the information, language, qualified personal and financial issues are some of the main constraints to use such improved methods. For example, in some Central American countries, the World Bank had to finance the reconstructions of climate variables using interpolation methods to promote weather index-based insurance for agriculture [20]. We consider that the proposed method

could be a practical alternative as it is based on information that could be available, and it does not require high computational resources or technical expertise.

3 Methodology

Variables experimentally observed or produced by reanalysis or modeling processes retain consistency among them. In our approach, we exploited this consistency among variables (correlated behavior) to build and parameterize a Bayesian network model, then inferring the missing values for a particular variable. It assumed that even if there is not a direct physical causal relationship among variables, some variables can work as proxies of others. The proposed methodology is summarized in Figure 1.

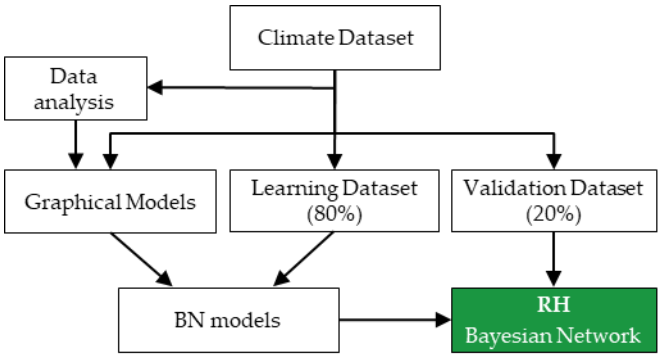

Fig.1: Model development process.

This study is part of a set of studies that explore the changes in the land suitability for coffee production in Central America. In particular, we are interested in modeling the variable Relative Humidity of the Driest Month (RHDM) for the region. The driest month refers to the month with the lowest precipitation. RHDM has been identified as an agro-ecological variable that influences the suitability of a site for coffee production [5] and is a common unobserved variable for meteorological stations.

3.1 Variable Selection

We used the surface reanalysis dataset: Climate Forecast System Reanalysis (CFSR) [7]. CFSR includes the variables precipitation, temperature (minimum

and maximum), wind speed, solar radiation and relative humidity (https://globalweather.tamu.edu). The spatial resolution is 38km per pixel at a daily time step for the period from 1979 to 2014. A complete dataset was downloaded for the extent of Central America. From this dataset, a sub-dataset for the relative humidity of the driest month was created (RHDM-dataset) by averaging daily to monthly values from multiples years to a single year, then extracting the variables values for the two driest months ($n = 1710$). Once the RHDM-dataset was built, we selected those variables with a significant correlation to RH, as depicted in Table 1. Maximum temperature ($Tmax$), minimum temperature ($Tmin$) and precipitation ($Rain$) were selected as proxy variables. A sensitivity analysis to findings was done to display the variance reduction of proxy variables on the relative humidity (RH).

According to the Pearson's correlation coefficients, all variables have significant relations to RH: $Tmax$ and $Rain$ have the major correlation, then solar radiation ($Solar$), and then $Tmin$ and wind speed ($Wind$) (see Table 1). However, in practice $Solar$ and $Wind$ are hardly measured so we excluded them.

	RH	Tmax	Tmin	Rain	Wind	Solar
RH	1.00	–	–	–	–	–
Tmax	-0.60	1.00	–	–	–	–
Tmin	0.12	0.12	1.00	–	–	–
Rain	0.60	-0.21	0.17	1.00	–	–
Wind	0.11	-0.30	0.65	-0.09	1.00	–
Solar	-0.15	0.19	0.67	-0.13	0.57	1.00

Table 1: Pearson's correlation coefficients between the climate variables for Central America.

Thus, the developed model uses only $Rain$, $Tmax$ and $Tmin$ as proxy-predictor variables for RH.

3.2 Model Development

The model was built using the software package Netica (www.norsys.com). For each variable, nodes were created, and descriptive statistics used to define the number and size of the states of the nodes. Proxy variables were linked directly to RH as parents; then a *link reversal*[1] proceeding was conducted to incorporate the correlation among the proxy variables (climate variable

[1] Reversal links: https://www.norsys.com/WebHelp/NETICA.htm

consistency). The priors and Conditional Probabilistic Tables (CPTs) were learned from the training dataset using the Counting- Learning Algorithm [14]. The model –structure, and parameters– is illustrated in Figure 2.

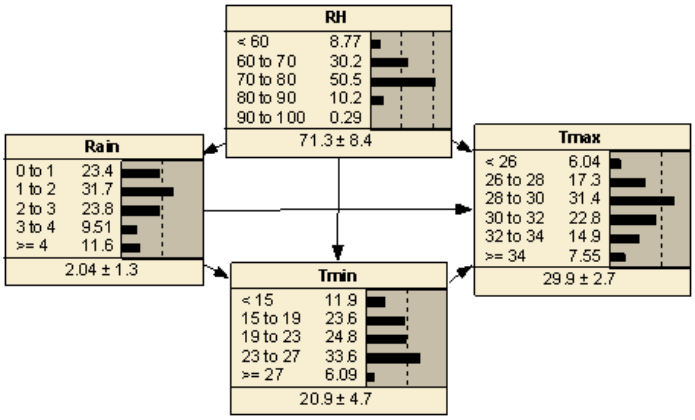

Fig. 2: Bayesian network for the relative humidity of the driest month. *RH*: relative humidity (%), *Tmax*: mean maximum temperature (°C), *Tmin*: mean minimum temperature (°C) and *Rain*: total precipitation (mm). Links do not represent biophysical relationships. Priors and conditional probabilities were learned from dataset CFSR (http://globalweather.tamu.edu).

4 Experiments and Results

4.1 Sensitivity analysis

The sensitivity analysis, Figure 3, agreed with the variable relevance obtained with the Pearson correlation analysis (Table 1). *Rain* and *Tmax* have the highest influence on *RH* and *Tmin* the lowest one.

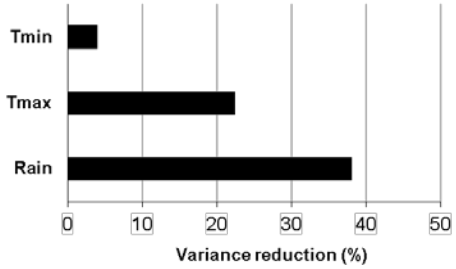

Fig. 3: Sensitivity analysis results for relative humidity. *Rain*: precipitation (mm), *Tmax*: maximum temperature, *Tmin*: minimum temperature.

4.2 Model Validation

The HRDM dataset was divided into 80% (n=1368) for training and 20% (n=342) for validation of the model. Using the validation dataset, we evaluated the model using different metrics. The Bayesian metrics logarithmic loss, quadratic loss, and spherical payoff were estimated using the feature *test with cases* in Netica [14]. The metrics RMSE, R2 (Adjusted), Index of Agreement and bias were calculated from the *expected RH* vs. *RH* from the validation dataset [2,11]. Also, it was graphically examined if the estimated RH maintains the relationship with the precipitation and temperature variables in the same magnitude as the RH of the original dataset –CFSR.

The results are summarized in Table 2. The Bayesian metric spherical pay- off indicates a good performance of the model for inferring RH [11]. Also, the more traditional metrics in meteorological sciences: bias close to zero, a very low RMSE, an index of agreement (d_2) over 0.90 and an R^2 of about 0.80 attested the very good performance of the model.

Metric	Values
Spherical	0.80
Bias	0.50
RMSE	3.64
d_2	0.93
R^2 (Adj)	0.79

Table 2: Performance metrics. Bayesian scoring metrics were estimated using the test with cases option in Netica. The traditional metrics were estimated using the expected value (mean value) of RH from the model vs. reanalysis values.

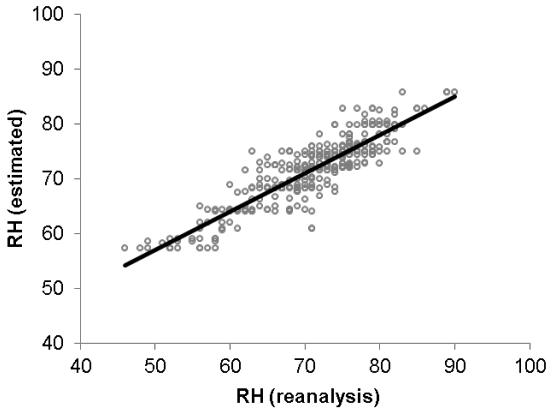

Fig. 4: Scatter plot of monthly relative humidity estimated vs. reanalysis (validation data set).

Figure 4 displays graphically the correspondence between the *RH* estimated for the model versus the reanalysis of the data reported in CFSR, showing, in general, a high consistency between the estimated and real (reanalysis) values.

We also confirmed that the estimated *RH* maintains consistency with the other variables as in the training data. This means that the estimated *RH* keeps the same relationship with the other variables, which is one of the major concerns in modeling climate data: consistency among data. This can be appreciated in Figure 5 which depicts scatter plots of *RH* vs. the other three variables, using the training (reanalysis) and estimated data.

Finally, a spatial comparison of the estimated vs. the actual (reanalysis) *RH* over the region (with a pixel size of 38 km) is depicted in Figure 6. We can appreciate that in general, the model reproduces very well the patterns of *RH* in the region.

4.3 Model application

The current state of the model allows using it in two ways to infer the relative humidity of the driest month (RHDM) for coffee production or the monthly RH. The model was created, adjusted and validated to generate the RHDM in Central America, where ranges between 50-60% are optimal and 80-90% suboptimal for Coffea arabica [5]. So, for land evaluation for coffee production, the model can

be used at two levels: first, at local or farm level, by using directly the model to estimate the RHDM; and second, at regional or national level, by using RHDM-suitability maps (created from data in Fig. 6: expected RHDM) to determine the optimal areas to implement coffee policies or programs. An example is displayed in Figure 7.

Inferring monthly *RH*. Even if the model was developed to infer *RHDM*, the model can estimate with an acceptable accuracy the monthly *RH* (values of any month). To do it, the CPTs were updated using the sub-dataset including all the months for the study region. Then the expected *RH* was estimated and compared with the reanalysis values. The model was capable of maintaining a very good performance, see Table 3. The results indicate that the model only required updating the CPTs using the monthly values for *RH*.

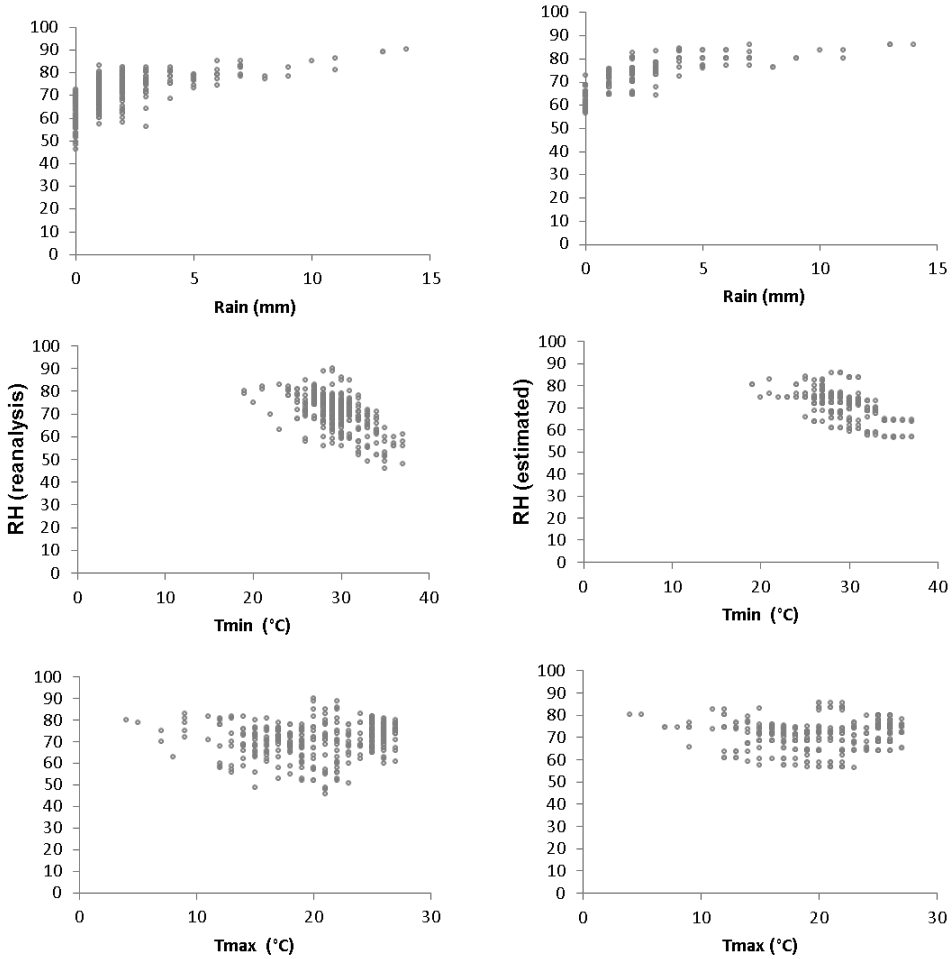

Fig. 5: Relative humidity of the driest month vs. proxy variables: left, reanalysis *RH*; right, estimated *RH*. Notice that the estimated *RH* against the proxy variables describes similar patterns than using *RH* from reanalysis. *RH*: relative humidity, *Tmax*: mean maximum temperature, *Tmin*: mean minimum temperature, and *Rain*: total precipitation. Monthly data.

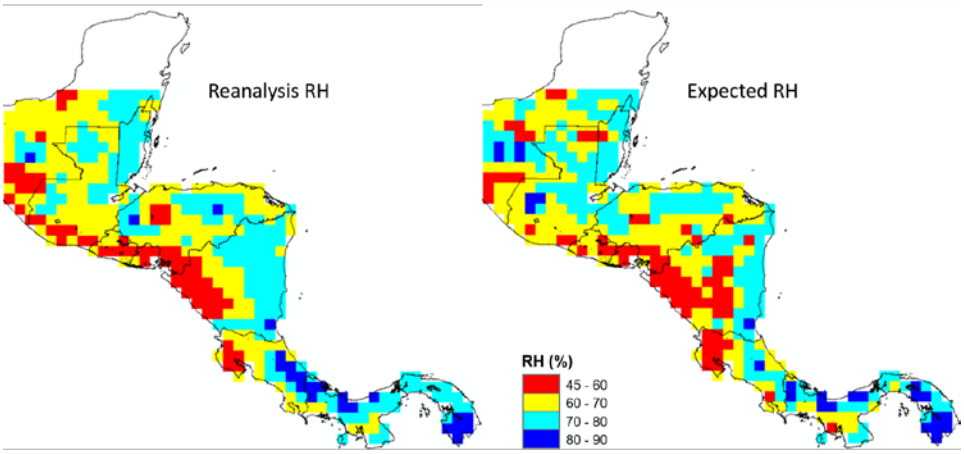

Fig. 6: Relative humidity of the driest month from reanalysis and estimated by the BN model for Southern Mexico and Central America (pixel size 38 km).

Metric	Values
Bias	0.48
RMSE	4.39
d_2	0.93
R^2	0.76

Table 3: Performance metrics for RH. The metrics were estimated using the expected value (mean value) of RH from the model vs. reanalysis values. d_2 =Index of agreement (0 to 1, where 1 is a perfect match).

Considerations for further applications. The model application to different regions or high-resolution data is possible following the same methodology, but the variable states should be revised according to the new conditions. In the case of using high-resolution data, we will require reviewing if the correlation among variables remains as we saw at gross resolution if this is not the case the model variables should be updated. At this point, the time step analysis is monthly values; further analysis have to be done to explore if the proxy variables and their states remain unchanged at finer time steps such as weekly or daily. Lastly, we consider that this method can be applied to overcome missing data for other climate variables.

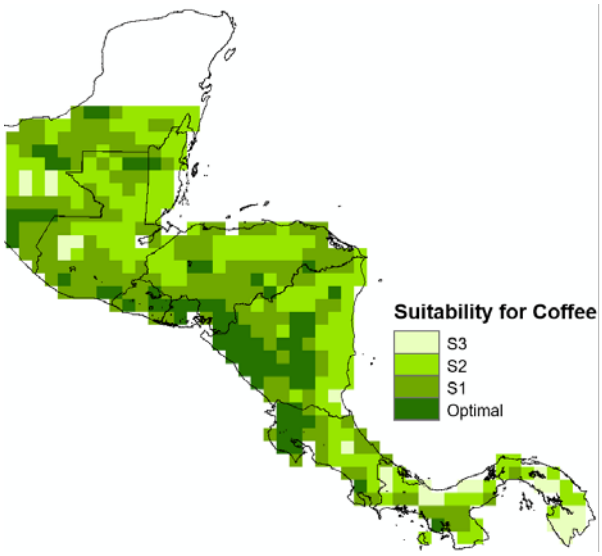

Fig. 7: Suitability map of Relative Humidity of the Driest Month for Coffea arabica L. for Southern Mexico and Central America (pixel size 38 km). Modified from [5]. Optimal = optimal conditions, S1 = Very good, S2 =Moderate, S3 = Marginal.

5 Conclusions and Future Work

We introduced a meteorological application of Bayesian networks to generate missing data of a climate variable. The procedure is simple, requiring a low modeling effort, and ensures maintaining the corresponding inter-relationship (consistency) to the others climate variables. The use of BN allows knowing the uncertainty of the inferred data, which is missing when traditional methods based on equations are implemented to estimate climate data.

We used as a case study the relative humidity (*RH*) of the driest month, which is one of many variable-indicators to describe the suitability of the land for coffee production. After few adjustments, the model could estimate monthly *RH* for all the year, in different regions or fine spatial resolution. Because *RH* is frequently unregistered in comparison to precipitation and temperature; we believe that our model offers to the region a valuable tool to generate reliable data for *RH*.

In the future, we plan to integrate this information into a Bayesian network model for analyzing land suitability for coffee production incorporating predictions of future climate changes.

References

1. Acock, M.C., Pachepsky, Y.A., 2000. Estimating missing weather data for agricultural simulations using group method of data handling. Journal of Applied meteorology 39, 1176?1184.
2. Badescu, V., 1996. Use of Willmott?s index of agreement to the validation of meteorological models. The Meteorological Magazine 122, 282?286.
3. Cofiño, A., Cano, R., Sordo, C., Gutierrez, J., 2002. Bayesian Networks for ProbabilisticWeather Prediction.
4. De la Torre-Gea, G., Soto-Zarazúa, G.M., Guevara-González, R., Rico-García, E., 2011. Bayesian networks for defining relationships among climate factors. Int. J. Phys. Sci 6, 4412?4418.
5. Descroix, F., Snoeck, J., 2004. Enviromental Factors Suitable for Coffee Cultivation, in: Wintgens, J. (Ed.), Coffe: Growing, Processing, Sustainable Production. Wiley-VCH, Alemania, pp. 164?177.
6. Frolking, S., Qiu, J., Boles, S., Xiao, X., Liu, J., Zhuang, Y., Li, C., Qin, X., 2002. Combining remote sensing and ground census data to develop new maps of the distribution of rice agriculture in China: PADDY RICE CROPLAND MAPS FOR CHINA. Global Biogeochemical Cycles 16, 38-1-38?10. doi:10.1029/2001GB001425
7. Fuka, D.R., Walter, M.T., MacAlister, C., Degaetano, A.T., Steenhuis, T.S., Easton, Z.M., 2014. Using the Climate Forecast System Reanalysis as weather input data for watershed models: USING CFSR AS WEATHER INPUT DATA FOR WATERSHED MODELS. Hydrological Processes 28, 5613?5623. doi:10.1002/hyp.10073
8. Ibargüengoytia, P.H., Vadera, S., Sucar L.E.: A Probabilistic Model for Information Validation. British Computer Journal. 49(1), 113–126 (2006)
9. Little, R.J., Schenker, N., 1995. Missing data, in: Handbook of Statistical Modeling for the Social and Behavioral Sciences. Springer, pp. 39?75.
10. Liu, D.L., Scott, B.J., 2001. Estimation of solar radiation in Australia from rainfall and temperature observations. Agricultural and Forest Meteorology 106, 41?59.
11. Marcot, B.G., 2012. Metrics for evaluating performance and uncer- tainty of Bayesian network models. Ecological Modelling 230, 50?62. doi:10.1016/j.ecolmodel.2012.01.013
12. Miller, D., Robbins, M., Habiger, J., 2010. Examining the challenges of missing

data analysis in phase three of the agricultural resource management survey. JSM Proceedings, Section on Survey Research Methods, Alexandria, VA: American Statistical Association.

13. Moxey, A., 1999. Cross-Cutting issues in developing agri-environmental indicators, in: Environmental Indicators for Agriculture: Issues and Design. Organization for Economic Co-operation and Development, Paris, France?: Washington, DC.
14. Norsys, 2015. Netica Help [WWW Document]. URL http://www.norsys.com/WebHelp/NETICA.htm (accessed 9.15.15).
15. Schneider, Tacock., 2001. Analysis of incomplete climate data: Estimation of mean val- ues and covariance matrices and imputation of missing values. Journal of Climate 14, 853?871.
16. Smith, T.M., Reynolds, R.W., Livezey, R.E., Stokes, D.C., 1996. Reconstruction of Historical Sea Surface Temperatures Using Empirical Orthogonal Functions. J. Climate 9, 1403?1420. doi:10.1175/1520-0442(1996)009<1403:ROHSST>2.0.CO;2
17. Sucar, L.E., 2015. Probabilistic Graphical Models: Principles and Applications. Springer-Verlag, London.
18. van Oijen, M., Dauzat, J., Harmand, J.-M., Lawson, G., Vaast, P., 2010. Coffee agroforestry systems in Central America: II. Development of a simple process-based model and preliminary results. Agroforestry Systems 80, 361?378. doi:10.1007/s10457-010-9291-1
19. van Wart, J., Grassini, P., Cassman, K.G., 2013. Impact of derived global weather data on simulated crop yields. Global Change Biology 19, 3822?3834. doi:10.1111/gcb.12302
20. World Bank, 2013. Weather data grids for agriculture risk management?: the case of Honduras and Guatemala (No. 77780). The World Bank.