SUBMITTED TO JOURNAL OF HYDROLOGY (JUNE 2010).

Downscaling dense rain gauge networks with short historical records using Bayesian networks

R. Ancell

Agencia Estatal de Meteorología (AEMET), Santander, Spain

J.M. Gutiérrez

Instituto de Física de Cantabria (CSIC-UC), Santander, Spain

ABSTRACT

Statistical downscaling is a mature field with many applications in different socioeconomic sectors —such as hydrology— which require local weather data. Standard downscaling methods rely on long historical records (tens of years) typically associated with sparse networks of low spatial resolution (e.g. the SYNOP network). Thus, these methods are not suitable for modern station networks with shorter temporal coverage (years) and, typically, higher spatial resolution. In this paper, we present a new technique to combine both sources of information into a single downscaling model, connecting the modern to the historical stations in a hierarchical way, thus avoiding overfiting. To this aim we use Bayesian networks to graphically represent the variables and their dependencies defining an underlying joint probabilistic model; we introduce a new automatic learning algorithm method to infer the model with an appropriate hierarchy from data. The resulting model is shown to improve the results of other state of the art methods, such as the Naïve model or a simple nearest neighbors method. In particular we consider precipitation occurrence and focus on probabilistic weather-typing downscaling techniques.

1. Introduction

During the last two decades a variety of statistical downscaling methods have been introduced in the literature (see, e.g. Wilby and Wigley 1997; Benestad et al. 2008, and references therein). Statistical downscaling methods work by projecting the upper-air large-scale circulation fields predicted by some Global Circulation Model (GCM) down to some surface local variable of interest; among these variables, precipitation is still one of the most challenging due to its mixed nature (occurrence and amount) and its non-gaussian character. Thus, many downscaling methods have been tuned and adapted to donwscale precipitation occurrence, providing a probability for this binary event. Some of these methods treat individually each particular station (single-site), and others work collectively with all the available stations (multi-site). In the later case, the spatial coherence of inter-station correlation must be preserved in the downscaled, thus imposing additional constraints to the downscaling methods (see, e.g., Harpham and Wilby 2005, and references therein). Multi-site methods are needed in hydrological applications in order to obtain the right aggregated values, e.g. the appropriate runoff amounts.

The enlargement of terrestrial observation systems with modern networks —typically with high spatial resolution— (Zepeda-Arce et al. 2000; Janis et al. 2004) poses new challenges for statistical downscaling, since standard methods rely on historical networks with long records (tens of years). Among the problems to be avoided, overfiting and loss of generalization play a central role, since the model parameters must be trained with only a few data (e.g. for a weather-type conditioned regression model) and the inference made on the basis of a few instances may not generalize to other situations (e.g. for the analog or weather typing approaches).

In this paper we present a new weather-typing technique to overcome these problems by combining both modern and historical networks into a single probabilistic downscaling model. The key idea is connecting all the available information (the modern stations, the historical ones and the weather types) in a hierarchical way. In the resulting model, connections from the weather type variable will be typically established with historical stations (based on long time records) and, in turn, historical stations will be connected with modern ones, thus indirectly transferring the information from the weather type; as we shall see later, the different parameters of the model correspond to local probabilities (from a reduced set of stations) and, thus, can be trained with the different time-slices of the available data in an optimum way avoiding overfiting. To this aim we use Bayesian networks to graphically represent the variables and their hierarchical dependencies, defining an underlying joint probabilistic model which can be used for probabilistic inference (e.g. downscaling, when providing the weather type as evidence). We introduce a new automatic learning algorithm to infer the model from data with an appropriate hierarchy, building on existing learning algorithms and taking advantage of the spatial character of the problem (see, e.g. Friedman et al. 1999; Cano et al. 2004, for an introduction of statistical learning in this context). The resulting model is shown to improve the results of other state of the art methods, such as the Naïve model or a simple nearest neighbors method.

The paper is organized as follows. First, in Sec. 2, we describe the area of the study and data used in this work. In Section 3 we introduce the probabilistic network models (in particular Bayesian networks) as a suitable data mining tool for meteorological applications, focusing on the probabilistic downscaling problem. Then, in Sec. 4 we present the proposed downscaling method (the H2 downscaling algorithm) and give some comparison results with other benchmark methods. Finally, in sec. 5, some conclusions are reported.

2. Area of Study and Data

The Cantabric coast is a region of complex orography and Atlantic climate located in Northern Spain (see Fig. ba). In this region we consider a total of 42 quality controlled rain gauges (see Table 1) maintained by the Spanish Met Service (Agencia Estatal de Meteorología, AEMET) covering different time-slices of the period 1957-2002 (the ERA40 reanalysis period). On the one hand the twelve stations shown as crosses in Fig. bc are historical stations with a common observation record of over 25 years (9860 days) with no missing data. On the other hand, the thirty stations marked with open circles in Fig. bc correspond to recent stations with a shorter record (4816 days, corresponding to 13 years). We analyze binary events associated with two different precipitation thresholds $precip > 0mm$ and $precip > 10mm$ (the frequency of these events for the different stations is shown in Table 1).

In the different experiments performed in the paper, a test period of approximately 4 years (1445 days) was randomly selected over the common 13 years period and removed from the data used to train the different models, keeping 8415 days (21 years) and 3371 days (9 years) for training the historical and recent stations, respectively. In the later case, as an additional experiment, the training period has been reduced to 337 days (1 year) in order to analyze the overfiting effects that may arise when using very recent stations. Thus, we shall explore the limits of the proposed methodology to include new stations (with data between 9 and 1 years) in the downscaling process.

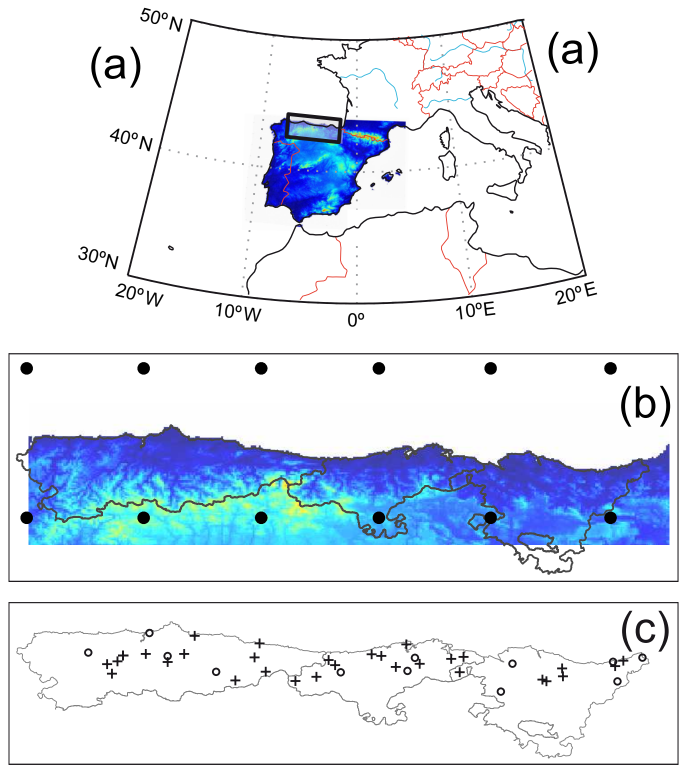

Fig. 1. (a) Region of study in Nothern Iberian Peninsula: the Cantabric coast (inner square). (b) Orography of the region and ERA40 reanalysis grid points (black solid points). (c) Historical (crosses) and recent (open circles) stations used in the study.

In order to define the atmospheric state to be used as predictor for the downscaling process, we use data from the Reanalysis ERA40 project (Uppala and others 2005), from the European Center for Medium-Range Weather Forecast (ECMWF). Following the indications of (Guti´rrez et al. 2004) we considered the meso-scale domain defined by the 12 black solid points in Fig. bb, in three levels 1000, 850 and 500 hPa. For each day of the ERA40 period (1957-2002), the corresponding atmospheric pattern is formed by the temperature, geopotential, North and East wind components and humidity fields, resulting a vector with 150 components. In this paper we consider a weather typing approach so we applied the *k-means* clustering method to the daily atmospheric vectors in the reanalysis space, obtaining a single discrete scalar weather type C^k, with k different states (see Guti´rrez et al. 2004, for more details). In order to analyze the sensitivity of the downscaling results to the resolution of the weather types (number of weather types), we have considered both a low resolution weather type with only $k = 10$ classes, and a mid resolution one with $k = 100$ classes (higher resolution weather types with $k > 100$ classes have been for short-range downscaling, but the series available in this work are not long enough to properly fit the resulting models).

3. Bayesian Networks

Bayesian networks (BNs) is a popular, sound and intuitive methodology which allows building probabilistic models from data in problems with a large number of variables (Castillo et al. 1997). This is done by considering only the relevant dependencies or associations among the variables. The basic idea of BNs is to encode these dependencies using a graphical representation (a directed acyclic graph) which is easy to understand and interpret (see Fig. 2 for some examples). The graph includes a node for each variable (in this case, the different stations and an extra variable for the weather-type representing the state of the atmosphere), and edges connecting variables which are directly dependent on each other. Every graph defines a decomposition of the high-dimensional Joint Probability Distribution (JPD) into a product of low-dimensional local distributions (marginal or conditional) as follows:

$$P(y_1, ..., y_n) = \prod_{i=1}^{n} P(y_i|\pi_i), \tag{1}$$

where each Y_i is a variable in the model, and Π_i are the parents of Y_i, i.e. the subset of variables whose edges are pointing directly to Y_i. Therefore, the independencies from the graph are easily translated into the probabilistic model in a sound and intuitive form. In this paper we shall consider $\mathbf{Y} = \{X_1, \ldots, X_n, C\}$, where X_i represent precipitation occurrence at the different stations and C is the weather-type variable defined in Sec. 2.

Learning Bayesian networks from data is a hard task, since the number of different models is exponential in the number of variables (Neapolitan 2003); therefore, different learning algorithms have been proposed in the literature to solve this problem from different points of view. One of the most popular approaches is the so-called *search-and-score*, where a search algorithm (in the space of feasible graphs) is used in conjunction with a scoring criterion to evaluate the fit of candidate models to the available data. When a good graph is identified, the parameters of the corresponding JPD in (1) are estimated from data in a straightforward way. Two popular search methods are the *K2* and *B* greedy algorithms, and a common score metric is the Minimum Description Length (MDL), which considers the likelihood of the data and a penalty term for model complexity, thus benefiting models with a simpler description of the data (see Neapolitan 2003, for details). In this paper we consider an alternative goal-oriented score metric, the ROC Skill Area (Jolliffe and Stephenson 2003), which allows evaluating the skill of the resulting model to predict the precipitation occurrence, $p(y_i|c)$, based on the weather-type variable state. This score metric has been already used to learn classifiers from data.

Different applications of Bayesian networks have been presented in the literature; for instance Abramson et al. (1996) describes an application for hail forecasting, Kennett et al. (2001) for sea breeze forecasting, and Hruschka and Ebecken (2005) for fog prediction. These studies are focused on a single site and consider the local forecast as a unidimensional problem. However, the collective behaviour of different sites is a key factor in hydrological applications, where the prediction for total basin precipitation is an aggregation of the different local predictions. In a recent work Cano et al. (2004) have shown that Bayesian networks are also sound and convenient models for multi-site standard applications for downscaling and local prediction, providing probabilistic models with spatial consistency able to simulate the collective behaviour of different stations. In this paper we consider the multi-dimensional focus of this problem, introducing the BN as a joint model to forecast all the locations preserving the spatial consistency at the same time.

a. *Downscaling with Bayesian Networks*

Following the work by Cano et al. (2004) we apply Bayesian networks to build a multivariate model including the dependencies among precipitation occurrence for the set of stations X_i shown in Fig. b. Moreover, we also consider the weather-type C described in Sec. 2 as an additional variable, to be used as evidence (known value) for downscaling purposes. Therefore, depending on type of graph (dependency model) considered, the Bayesian network can be seen as a particular multi-site weather-typing probabilistic downscaling technique. The simplest Bayesian network structure is the Naïve Bayes, where the weather-type variable is the only parent of all the stations, as shown in Fig. 2(a). This model is based on the assumption that the stations $X_1, ..., X_n$, are conditionally independent given the weather-type value C:

$$I(X_i, X_j|C)\ \forall i, j \tag{2}$$

Thus, the corresponding JPD can be factorized as:

$$P(x_1, ..., x_n, c) = P(c) \prod_{i=1}^{n} P(x_i \mid c). \tag{3}$$

Therefore, the performance of the Naïve model for a particular application will depend on whether the different variables meet the assumptions given in (2). Note that this can be tested by considering, e.g., the conditional mutual information of each pair of stations for each particular weather-type state: $MI(X_i, X_j|C)$. Figure 3 illustrates behavior of the Naïve model for the case analyzed in this paper, considering the stations shown in Fig. b(c). The value of the conditional mutual information for each pair

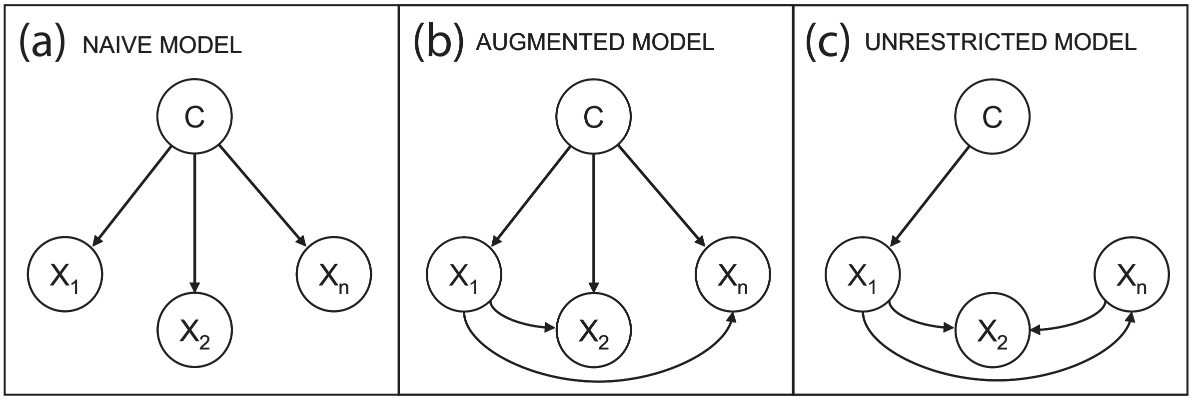

FIG. 2. Examples of three different directed acyclic graphs with a weather-type variable C and n stations X_i corresponding to: (a) A Naïve model with no direct dependencies among the stations, (b) an augmented model, obtained including additional links to the Naïve model, and (c) an unrestricted model with arbitrary dependencies.

of stations is plotted versus the distance between them, represented as $X_i||X_j$. This figure shows an exponential decay of the mutual information as time increases; thus, the the Naïve model is only appropriate in sparse networks with long inter-station distances (200 km in this example, if we consider a threshold of 0.1 for the mutual information). According to this, the figure has been divided into four areas depending on whether $MI(X_i, X_j|C) > 0.1$ and $X_i||X_j > 200km$. Thus, in the upper left area the stations are conditional dependency (labelled with crosses), whereas in the lower right area they are conditional independency (labelled with solid points). Therefore, the Naïve model could be appropriate for the historical stations (with low spatial resolution), but not for the modern stations with a higher resolution and small inter-site distances.

Since the forecast skill of the Naïve model relies only on the weather type variable, we can conclude that the Naïve model hypothesis is appropriate only in cases where the atmospheric model has higher spatial resolution than the observations; for instance, this is what happens with low resolution observation networks (which corresponds with solid points in Fig. 3). However, if the observed variability is higher than the atmospheric model resolution, then (2) is not meet and, thus, the Naïve model does not properly represent all the relevant dependencies among stations (see crosses in Fig. 3).

In spite of its simplicity, the Naïve model have shown a good performance compared with more complex models (see, e.g. Zhang et al. 2005). Therefore, the structure of this model has been taken as the basis to build Augmented Naïve models taking into account the spatial dependency among stations (see e.g. Fig. 2b). Examples of this are the *Selective Bayesian Classifiers* (SBC) proposed by Langley and Sage (1994), the *Tree Augmented Bayesian Classsifers* (TAN) proposed by Friedman et al. (1997), the *Aggregating One-Dependence Estimators* (AODE) from Webb et al. (2005), or Langseth and Nielsen (2006), wich uses Hierarchical Naïve classifers adding latent variables to the attributes, among others. Using Bayesian networks it is also possible to consider general unrestricted models where all the variables (stations and weather-type) are treated in equal conditions and the more relevant dependencies among them are considered (see Fig. 2c).

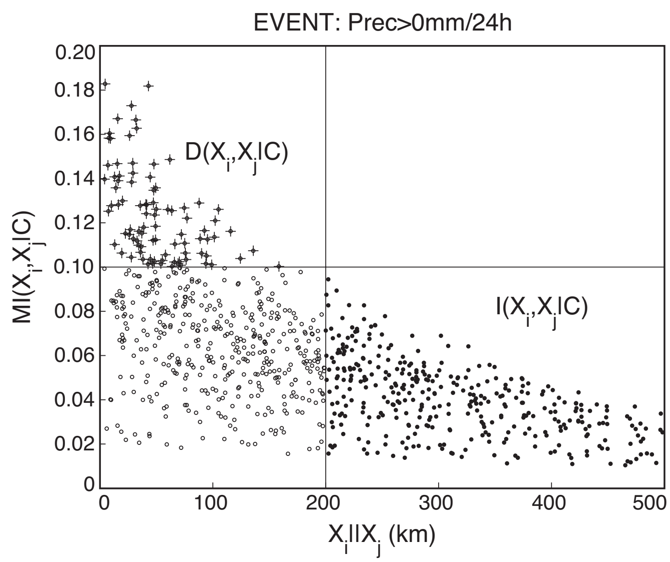

FIG. 3. Conditional mutual information between pairs of locations as a function of their distance for the occurrence of precipitation.

These models can be used in different ways depending on the evidence and target variables. On the one hand, they work as a classifier when the evidences are the stations $\mathbf{X}$ (or a particular subset $\mathbf{Z} \subset \mathbf{X}$), and the target is the weather-type variable C, so the goal is to estimate $p(c|\mathbf{z})$. On the other hand, they work as a downscaling model when the evidence is the weather-type C and the targets are the stations; in this case the goal is computing $p(\mathbf{z}|c)$ or, more simply, $p(x_i|c)$, for some i.

4. Downscaling Modern Networks with Short Historical Records

A common situation in many regional impact studies is the co-existence of sparse networks with long historical records (a few stations over the area of study) and a high-resolution network of newer stations with short historical records (typically one or two years). In this case, even the simplest Naïve model can be overfitted to data since the number of parameters in (3) is conditioned by the weather-type, which may have up to 100 states in practical applications. In this case, due to the heterogeneity of the available data, a generic structure with no pre-impossed dependencies (Fig. 2c) would lead to optimal models, penalizing those connections leading to a high number of parameters with poor prediction capabilities (e.g. links from the weather type to the newer stations).

The overfitting problem is illustrated in Fig. 4, which shows the validation scores over the test sample (described in sec. 2) for a Naïve model trained on the stations shown in Fig. b, but considering a common training period ranging from 1 to 10 years. In this case we can use standard validation techniques derived from 'hit' and 'false alarm' rates for the validation of the forecasts. Relative Operating Characteristic (ROC) curves and ROC Skill Areas (RSA) provide a global description of the skill of the method (see Jolliffe and Stephenson 2003, for an overview of validation for probabilistic forecasts).

The poor scores obtained up to 5 years in this case are due to the overfitting of parameters to the trainning data. Thus, in this example, we would need stations with at least 5 years of data in order to perform a robust statistical downscaling of the occurrence event with a Naïve model. Note that for more extreme events the period would increase (Jagannathan and Arlqy 1967). Therefore, to build models considering both historical and modern stations we need a simpler model build on the results of the Naïve hypothesis violation shown in Fig. 3. In the following section, we describe the $H2$ algorithm, a hierarchical algorithm where connections from the weather type variable will be typically established with historical stations (based on long time records) and, in turn, historical stations will be connected with modern ones, thus indirectly transferring the information from the weather type. The optimal structure for a particular application is inferred from data, thus providing an efficient downscaling methods that takes into account the particular restrictions imposed by the available dataset.

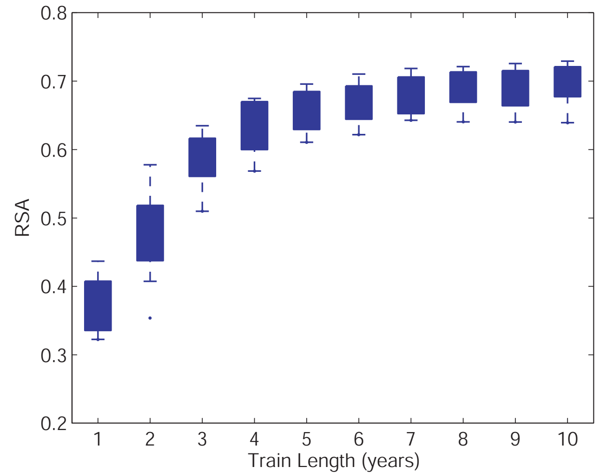

FIG. 4. Marginal quality (in RSA units) evolution of the Naïve model, showing overfitting effects for train lengths up to 10 years, using a C^{100} predictor for the $prec > 10mm/24h$ binary event. For every train length, each boxplot refers to 10 randomly selected train/test sets.

a. The H2 Algorithm

Let us consider a Naïve model, as defined in Fig. 2a with the binary variables $\mathbf{X}$ and the 100 state weather-type variable C^{100}, and let us divide the stations in two sets $\mathbf{X} = \{\mathbf{X_\Omega}, \mathbf{X_\Phi}\}$ corresponding to the long historical stations and the newer ones with short records, respectively, as described in Sec. 2. Note that each connection from C to X_i involves a large number of parameters due to the 100 different states of C (note that X_i are binary variables corresponding to the occurrence of precipitation in station i), whereas connections of the type X_i to X_j involve just two, since both variables are binary. In principle, the first connections would be only suitable for historical stations with long records, whereas the last one would be appropriate to connect new to historical stations; in this way older gauges would act as 'decoders' of the weather-types for the new ones. In this way the resulting model would avoid overfitting, combining all the heterogeneous available information into a single global model, allowing us to obtain downscaled values in a general framework. This is the idea that motivates the creation of H2 algorithm described as follows:

1 The initial network B is a *Naïve* model, formed by C as the only parent of every $X_i \in \mathbf{X_\Omega}$; note that the modern stations $\mathbf{X_\Phi}$ are disconnected.

2 Iterate 3-5 until no quality improvement is achieved:

3 Calculate the set of "candidate links", as those from nodes with at least one parent to stations in $\mathbf{X_\Phi}$ without children and no more than two parents (this constraint on the maximum number of parents could be relaxed, or even removed, depending on the application).

4 For every candidate link, evaluate the quality of the network B' obtained by adding the link to the current network B using $RSA(X_i) = f(P_{B'}(x_i|c))$; note that the marginal RSA value is used in this example, although other metrics can be used instead.

5 Update the current network B by including the candidate link leading to a maximum quality increment: $B \leftarrow B'$.

The limitation to a maximum of two parents have demonstrated to be be a good compromise between having a simple model (reducing the parameters) and a good performance. Note that links to modern stations which already have a child are forbidden because it means that they have been improved at least once, avoiding an unnecessary complex algorithm (a pseudo-code of the algorithm is included below). The name of the algorithm *H2* indicates the division in two different sets (historical and modern) to build the hierarchy (H) of the graph; note that a generalization of this algorithm to larger hierarchies is possible.

Algorithm 1 *H2* Structural learning algorithm.

Require: A sample of data, divided into three groups $\mathbf{D} = \{C, \mathbf{X_\Omega}, \mathbf{X_\Phi}\}$, where $\Omega = \{2, ..., m\}$ refers to the potential parents and $\Phi = \{m+1, ..., n\}$, refers to the potential sons, being C the weather type variable. The initial graph is defined by the adjacency matrix B: $B(1, j) = 1 \ \forall j \in \Omega$. The maximum number of parents per node is fixed as r.

while $\Phi \neq \emptyset$ **do**
  for $i \in \Phi$ **do**
    $S = score(\mathbf{D}, B); k = \emptyset;$
    for $j \in \Omega$ **do**
      if $acyclic(B \cup \{j \rightarrow i\}) = 1$ **then**
        $M = score(\mathbf{D}, B \cup \{j \rightarrow i\});$
        if $M > S$; **then** $S = M; k = j$; **endif**
      end if
    end for
    if $k \neq \emptyset$ **then**
      $B(k, i) = 1;$
      $\Omega = \{j \in [2, n] : \pi_j \neq \emptyset\};$
      $\Phi = \{i \in [m+1, n] : children(i) = \emptyset \wedge \sharp(\pi_i) \leq r\};$
    else
      $\Phi = \{\Phi \backslash \{i\}\};$
    end if
  end for
end while

Ensure: A directed acyclic graph B and his quality measure S.

Fig. 5 shows a graphical illustration of the **H2** algorithm. At each step of the algorithm there are three types of nodes: The weather type C, in black, the *predictors* (initially the historical stations), in gray, and the *predictands* (initially the modern stations) in white, which can receive links. There are also two types of links, the black ones with 100 parameters and the red ones with 2 parameters. In each iteration of the *H2* algorithm, $m \times (n - m)$ RSA different values must be computed, where m are the predictands with parent, and $(n - m)$ are the predictands without children and no more than two parents. Note that this number can be reduced using a local search criterion (see Friedman et al. 1999).

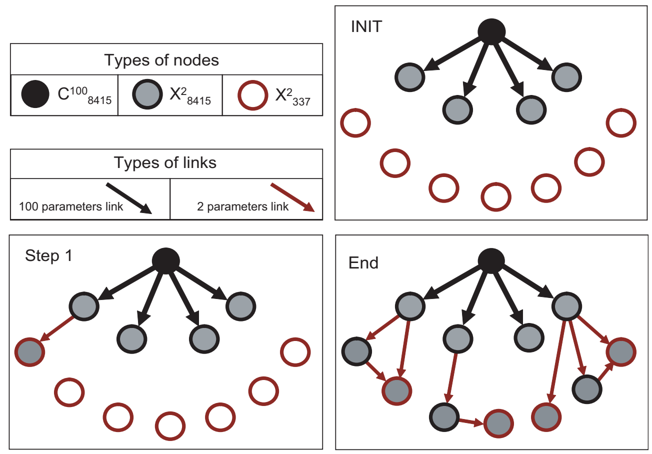

FIG. 5. Graphical illustration of the **H2** algorithm showing the initial state (INIT) and the available *predictors* at two different illustrative steps (after the first iteration and after the final one). The super and sub-indices of the types of nodes show the number of states and the size of the data available for the particular node, respectively.

Thus, in contrast to Naïve models, the resulting *H2* models explicitly include spatial relationships among the different stations which are not implicitly given by the weather type. Note that in the Naïve model the spatial consistency is only captured trough the weather type variable.

b. *Results and Comparison with Benchmark Methods*

In this section we describe the results obtained when applying the above algorithm to the data described in Sec. 2. To this aim, we consider the historical (crosses) and modern (open circles) stations shown in Fig. and apply the algorithm to the available data, considering different training subsets, from 1 to 10 years (the validation subset is kept constant in all the cases). The models are obtained using the training subset of the data, whereas the resulting models are validated using the test subset (the average of the RSA values of the different stations is used as the validation score). Fig. 6 shows the *H2* model obtained for the case *precip* $> 10mm$ and Figures 7(a) and (b) show the resulting RSA values for the case of a weather type with 10 and 100 states, respectively (red boxes); note that higher RSA values are obtained with a weather type with 100 possible states (100 different atmospheric configurations) and no overfitting is observed for those case with shorter training dataset. The results are similar for the *prec* $> 0mm/24h$ event, so they are not shown in the following.

With the aim of comparing the obtained results with other benchmark methods, we have also considered the Naïve model, and a nearest neighbors interpolation procedure. Downscaling interpolation models attempt to obtain estimates of probability at locations where predictions are not available, as a function of predictions available in other nearby locations (see Wilby and Wigley 1997, and references therein for more details). We have used a simple interpolation model that consists of estimating the probability of every *modern* station as the average of the predicted probabilities in ω_i, where ω_i is a subset of Ω with the three nearest historical stations; note that more complex interpolation models, considering different weights, could be considered but this is beyond the scope of this paper. The results for these two methods are also given in Figures 7(a) and (b) in blue and black colors, respectively. Note that the Naïve model has slightly better results when considering only 10 weather types. However, in the case of 100 weather types the Naïve model severely suffers from the overfitting problem and is only able to reach the performance of the *H2* algorithm in the case of long training series. In this case, the *H2* algorithm always outperforms the interpolation method.

Table 1 shows the individual RSA of the different stations obtained with the Naïve(N_1) and *H2* ($H2_1$) models for the first random realization of the case with 100 weather types and 1 year of training data (the case with worst possible overfitting). As expected, there are no changes in the 12 historical gauge stations; however, for the 30 modern stations, the model *H2* exhibits higher RSA values than the Naïve model. For the sake of comparison, a third column corresponding to the results with the long training series (10 years) is also shown under the label N_{10}. ESPINAMA location has been highlighted because improvements are not as good as in other locations, probably due to the very complex orography nearby (it is located in a very deep valley).

Finally, Fig. 8 shows the individual results for the 30 modern stations. It can be seen that although the interpolation and *H2* models have similar averaged performance (see Fig. 7), the situation is different when analyzing the results station by station. In particular, as shown in Fig. 8, there are some modern stations where *H2*is clearly better; these stations correspond to areas with high climatic variability, where the interpolation method cannot capture the local properties from the neighbors.

5. Conclusions

In this paper we have proved the potential of modern statistical learning techniques (Bayesian networks) to downscale heterogeneous networks with historical and modern gauges characterized by long and short historical records, respectively. Bayesian networks are able to globally model all the available information (stations and weather type) by inferring a joint probability distribution from data, taking into account the sparse dependency structure among stations. We analyzed the case of binary events (occurrence of precipitation considering two different thresholds), but this methodology can be generalized for the continuous case of precipitation amounts using, for instance, conditional-gamma Bayesian networks (see, e.g. Cowell 2006).

We have shown that Naïve models are the best models for individual probabilistic downscaling when the length of the series is long enough to prevent overfitting or when the observed resolution is lower than the modeled resolution (the one encoded within the weather types); in this case, the spatial dependency among the stations is implicitly captured from the weather type, yielding to the independence of each pair of stations given the weather type (Naive assumption). However, in cases with both historical and modern stations with short and long records and low and high spatial resolution, respectively, Naïve models are not appropriate and more elaborated models based on Bayesian networks can be learnt from data. In particular we present a hierarchical method (the *H2* algorithm) where the connections from the weather type variable are typically established with historical stations (based on long time records) and, in turn, historical stations are connected with modern ones, thus indirectly transferring the information from the weather type; we have shown that the different parameters of the model correspond to local probabilities (from a reduced set of stations) and, thus, they can be trained with the different time-slices of the available data in an optimum way avoiding overfiting.

These experimental results show that this model provides significantly better results than the Naïve and simple interpolation methods, concluding that the purposed methodology is skillful enough to allow incorporating modern stations into operational downscaling with only one year of available daily data.

Acknowledgement

The authors are grateful to the Spanish Weather Service (AEMET) for the data and partial support of this work.

REFERENCES

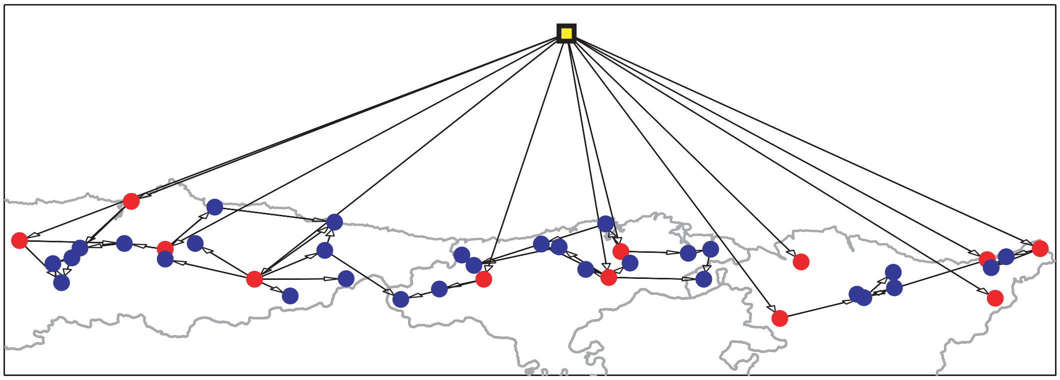

FIG. 6. Example of *H2* network for the $prec > 10mm/24h$ binary event, with the C^{100} model. Notice that the historical series (labelled in red) are the only ones with a direct link with the weather type (labelled with a yellow square), which acts as evidence.

Abramson, B., J. Brown, J. Edwards, A. Murphy, R. Winkler, and L. Robert, 1996: Hailfinder: A bayesian system for forecasting severe weather. **12 (1)**, 57–71.

Benestad, R. E., I. Hanssen-Baur, and D. Chen, 2008: *Empirical-statistical downscaling*. World Scientific Publishing.

Cano, R., C. Sordo, and J. M. Gutiérrez, 2004: Bayesian networks in meteorology. *Advances in Bayesian Networks*, J. A. Gámez, S. Moral, and A. Salmerón, Eds., Springer Verlag, 309–327.

Castillo, E., J. M. Gutiérrez, and A. S. Hadi, 1997: *Expert Systems and Probabilistic Network Models*. Springer-Verlag, free Spanish version `http://personales.unican.es/gutierjm`.

Cowell, R. G., 2006: A conditional-gamma bayesian network for dna mixture analyses. bayesian analysis. (to appear). a short synposis of this paper will appear in the proceedings of the valencia/isba eighth world meeting on bayesian statistics. *Bayesian Analysis*, **2**.

Friedman, N., D. Geiger, and M. Goldszmidt, 1997: Bayesian network classifiers. *Machine Learning*, **29 (2–3)**, 131–163.

Friedman, N., I. Nachman, and D. Peer, 1999: Learning bayesian network structure from massive datasets: The 'sparse candidate' algorithm. *Procedings of the 15th Conference on Uncertainty in Artificial Intelligence*, 206–215.

Gutiérrez, J. M., A. S. Cofiño, R. Cano, and M. A. Rodríguez, 2004: Clustering methods for statistical downscaling in short-range weather forecast. *Monthly Weather Review*, **132 (9)**, 2169–2183.

Harpham, C. and R. Wilby, 2005: Multi-site downscaling of heavy daily precipitation occurrence and amounts. **312**, 235–255.

Hruschka, E. and N. Ebecken, 2005: Applying bayesian networks for meteorological data mining. *Applications and Innovations in Intelligent Systems XIII*, Springer-Verlag, 122–133.

Jagannathan, P. and R. Arlqy, 1967: A note on climatological normals. Tech. Rep. 84-208, W.M.O.

Janis, M., K. Hubbard, and K. Redmond, 2004: Station density strategy for monitoring long-term climatic change in the contiguous united states. **17**, 151–162.

Jolliffe, I. and D. Stephenson, 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley and Sons.

Kennett, R., K. Korb, and A. Nicholson, 2001: Seabreeze prediction using bayesian networks. **2035**, 1611–3349.

Langley, P. and S. Sage, 1994: Induction of selective bayesian classifiers. *Procedings of Uncertainty in Artificial Intelligence*, Morgan Kaufmann, 400–406.

Langseth, H. and T. D. Nielsen, 2006: Classification using hierarchical naïve bayes models. **63 (2)**, 135–159, doi: http://dx.doi.org/10.1007/s10994-006-6136-2.

Neapolitan, R., 2003: *Learning Bayesian Networks*. Prentice-Hall, Inc.

Uppala, S. M. et al., 2005: The ERA-40 re-analysis. **131**, 2961–3012.

Webb, G. I., J. R. Boughton, and Z. Wang, 2005: Not so naive bayes: Aggregating one-dependence estimators. **58 (1)**, 5–24, doi:http://dx.doi.org/10.1007/s10994-005-4258-6.

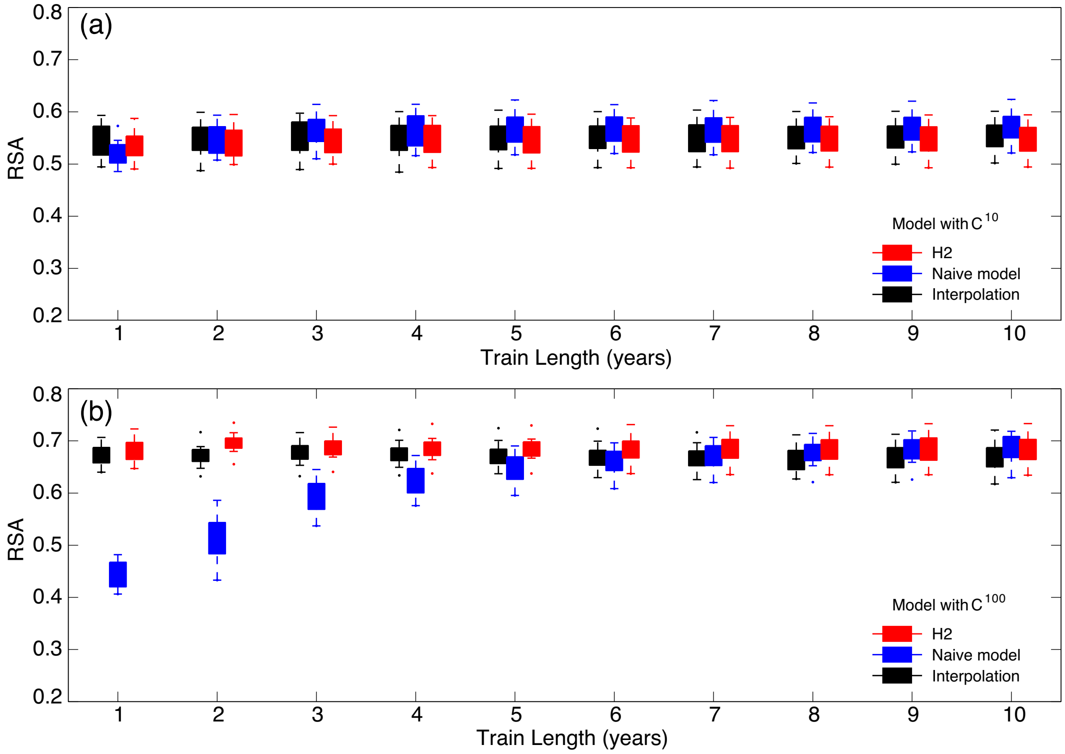

FIG. 7. RSA (average over all stations) corresponding to different train lengths, from 1 to 10 years for the $prec > 10mm/24h$ event with (a) 10 and (b) 100 weather types; the boxplots indicate the statistics of 10 different realizations of the experiment with different random training sets. In each case, the results of the *H2* method (red boxes) are compared with two benchmarks: the Naive model (blue) and a 3-neighbor interpolation method (black).

Wilby, R. L. and T. M. L. Wigley, 1997: Downscaling general circulation model output. A review of methods and limitations. *Progress in Physical Geography*, **21**, 530–548.

Zepeda-Arce, J., K. Foufoula-Georgiou, and K. Droegemeier, 2000: Space-time rainfall organization and its role in validating quantitative precipitation forecasts. **10**, 129–146.

Zhang, H., L. Jiang, and J. Su, 2005: Augmenting naive bayes for ranking. *ICML '05: Proceedings of the 22nd international conference on Machine learning*, ACM, New York, NY, USA, 1020–1027, doi:http://doi.acm.org/10.1145/1102351.1102480.

$p > 0$	$p > 10$	Location	N_1	$H2_1$	N_{10}
0.52	0.16	FUENTERRABIA 'AEROPUERTO'	0.80	0.80	0.77
0.58	0.15	SAN SEBASTIAN 'IGUELDO'	0.73	0.73	0.70
0.39	0.18	ELDUAYEN	0.75	0.75	0.70
0.32	0.10	AMURRIO 'INSTITUTO'	0.76	0.76	0.72
0.54	0.11	BILBAO 'AEROPUERTO'	0.80	0.80	0.76
0.49	0.14	EL MERCADILLO DE LIERGANES	0.75	0.75	0.71
0.41	0.14	VILLACARRIEDO	0.78	0.78	0.73
0.46	0.09	ROZADIO	0.79	0.79	0.77
0.53	0.10	RANON 'AEROPUERTO'	0.73	0.73	0.76
0.37	0.13	RIOSECO DE SOBRESCOBIO	0.75	0.75	0.81
0.54	0.08	OVIEDO 'EL CRISTO'	0.71	0.71	0.74
0.49	0.11	ZARDAIN	0.75	0.75	0.78
0.50	0.16	SAN SEBASTIAN 'ATEGORRIETA'	0.40	0.64	0.58
0.49	0.16	LASARTE 'MICHELIN'	0.42	0.68	0.69
0.50	0.14	EIBAR 'BANCO DE PRUEBAS'	0.40	0.78	0.73
0.56	0.14	ECHEVARRIA	0.46	0.78	0.73
0.44	0.14	ABADIANO 'MENDIOLA'	0.44	0.80	0.73
0.37	0.13	DURANGO 'VIVERO'	0.36	0.75	0.68
0.39	0.14	GURIEZO 'G.C.'	0.38	0.75	0.77
0.40	0.12	CARRANZA	0.46	0.75	0.76
0.49	0.13	COTERILLO DE AMPUERO	0.45	0.80	0.81
0.54	0.18	MIRONES	0.55	0.80	0.70
0.56	0.11	SANTANDER 'CENTRO'	0.33	0.72	0.72
0.44	0.13	ESCOBEDO DE VILLAFUFRE	0.32	0.69	0.64
0.47	0.11	TORRELAVEGA 'SNIACE'	0.31	0.79	0.77
0.55	0.10	CELIS	0.39	0.79	0.79
0.46	0.11	CAMIJANES	0.41	0.68	0.76
0.34	**0.09**	**ESPINAMA**	**0.33**	**0.30**	**0.66**
0.35	0.05	TAMA	0.46	0.61	0.81
0.42	0.16	AMIEVA	0.37	0.69	0.72
0.40	0.09	CANGAS DE ONIS	0.35	0.69	0.77
0.41	0.10	RIBADESELLA 'FARO'	0.39	0.71	0.77
0.52	0.08	GIJON	0.27	0.70	0.79
0.31	0.13	BEZANES	0.26	0.64	0.66
0.47	0.09	SOTO DE RIBERA	0.37	0.72	0.78
0.47	0.09	MERES DE SIERO	0.29	0.74	0.81
0.42	0.09	GRADO	0.39	0.73	0.73
0.38	0.13	GENESTAZA	0.41	0.68	0.73
0.37	0.08	SOTO DE LA BARCA	0.34	0.69	0.78
0.42	0.09	PRESA DE LA BARCA	0.43	0.69	0.73
0.48	0.11	SOTO DE LOS INFANTES	0.22	0.44	0.56
0.45	0.11	CUEVAS DE ALTAMIRA	0.28	0.77	0.68

Table 1. The first two columns of the table are the climatological occurence probabilities of $prec > 0mm/24h$ and $prec > 10mm/24h$, respectively; then, the name of the stations is shown. The last three columns show the $RSA(X_i)$ values of the first random realization for Naïve (N) and H2 model (the one corresponding to the graph of Fig. 6) corresponding to the case of one year training length, and the RSA of the Naïve model trained with 10 years of data.

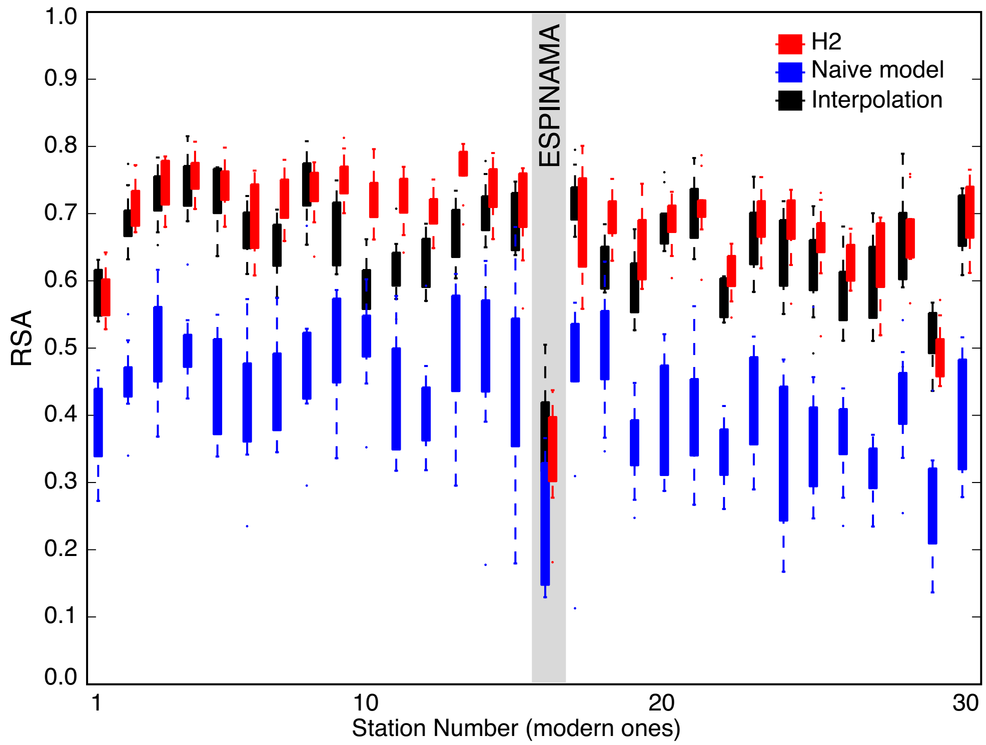

FIG. 8. Individual RSA values for each of the modern stations. Every boxplot corresponds to 10 random realizations with the short train set (1 year).