Causal Discovery for Climate Research Using Graphical Models

IMME EBERT-UPHOFF

Department of Electrical and Computer Engineering, Colorado State University, Fort Collins, Colorado

YI DENG

School of Earth and Atmospheric Sciences, Georgia Institute of Technology, Atlanta, Georgia

(Manuscript received 12 July 2011, in final form 4 February 2012)

ABSTRACT

Causal discovery seeks to recover cause–effect relationships from statistical data using graphical models. One goal of this paper is to provide an accessible introduction to causal discovery methods for climate scientists, with a focus on constraint-based structure learning. Second, in a detailed case study constraint-based structure learning is applied to derive hypotheses of causal relationships between four prominent modes of atmospheric low-frequency variability in boreal winter including the Western Pacific Oscillation (WPO), Eastern Pacific Oscillation (EPO), Pacific–North America (PNA) pattern, and North Atlantic Oscillation (NAO). The results are shown in the form of static and temporal independence graphs also known as *Bayesian Networks*. It is found that WPO and EPO are nearly indistinguishable from the cause–effect perspective as strong simultaneous coupling is identified between the two. In addition, changes in the state of EPO (NAO) may cause changes in the state of NAO (PNA) approximately 18 (3–6) days later. These results are not only consistent with previous findings on dynamical processes connecting different low-frequency modes (e.g., interaction between synoptic and low-frequency eddies) but also provide the basis for formulating new hypotheses regarding the time scale and temporal sequencing of dynamical processes responsible for these connections. Last, the authors propose to use structure learning for *climate networks*, which are currently based primarily on correlation analysis. While correlation-based climate networks focus on *similarity* between nodes, independence graphs would provide an alternative viewpoint by focusing on *information flow* in the network.

1. Introduction

One of the best known computational approaches to causality is the concept of Granger causality introduced by Granger (1969). A time series, X, Granger causes a second time series, Y, if past values of X contain information that helps predict future values of Y above and beyond the information contained in the past values of Y alone. Granger causality is implemented by first performing linear regression of the time series and then applying statistical tests on the regression coefficients. Granger causality is thus a measure for predictability based on a linear model and applies only to time series data.

Reasoning about causality was put on a more general footing starting in the late 1980s through the introduction of causal calculus (Rebane and Pearl 1987) and the use of probabilistic graphical models to represent causal relationships. The idea of representing causal structure in a graphical way goes back to Wright (1921, 1934) who defined path diagrams for structural equation models, a concept commonly used in economics to date. Pearl (1988) proposed the use of graphical models to represent probabilistic independence relationships between variables. This approach does not rely on temporal information, so it applies equally to nontemporal and time series data. Spirtes, Glymour, and Scheines (Spirtes et al. 1991, 1993) addressed the problem of detecting hidden common causes, which in turn allowed for causal interpretation of the graphs. These contributions by Pearl and Spirtes et al. laid the foundation for the field of causal discovery and thus jump started the development of a myriad of algorithms that detect cause–effect

Corresponding author address: Yi Deng, School of Earth and Atmospheric Sciences, Georgia Institute of Technology, 311 Ferst Drive, Atlanta, GA 30332-0340.
E-mail: yi.deng@eas.gatech.edu.

DOI: 10.1175/JCLI-D-11-00387.1

relationships from observational data (Spirtes et al. 2000; Pearl 2000; Neapolitan 2003; Koller and Friedman 2009). Even Granger later incorporated Pearl's graph approach, calculating graphs based on Granger causality tests for multivariate time series regression models (Swanson and Granger 1997; Eichler 2007). These models are also known as Graphical Granger models (Arnold et al. 2007).

The intent of this paper is to provide an introduction to causal discovery using graphical models for researchers in climate science and to demonstrate their use for an example in climate science. Causal discovery algorithms generate one or more graph representations that describe the potential causal pathways in the system. The most common type of graph used is a *Bayesian network* (Pearl 1988), which consists of two parts, a graph structure and probabilities, and all causal relationships are encoded in the graph structure. Causal discovery has already been applied with great success in disciplines ranging from the social sciences to computer science, engineering, medical diagnosis and bioinformatics (Spirtes et al. 2000; Neapolitan 2003). Many of the most successful examples in recent years come from the area of computational biology. For example, Margolin et al. (2006) and Friedman et al. (2000) trained Bayesian networks on expression data to identify protein/gene interaction, applying causal discovery to networks containing tens of thousands of nodes (Margolin et al. 2006).

In climate science, Bayesian networks have been primarily used for purposes such as forecasting or as risk assessment or decision-making tools, not to generate causal hypotheses. Since here we are more interested in learning potential causal relationships (i.e., graph structure of Bayesian networks) than quantifying probabilities, we categorize the following discussion of the relevant literature by the level of structure learning taking place. Work in the *first category* derives the structure of the Bayesian network directly from expert knowledge, and only probabilities are learned from data. A good example is the Hailfinder project by Abramson et al. (1996), which was one of the first applications of Bayesian networks related to climate science. Hailfinder is a Bayesian network for the prediction of severe weather events in northern Colorado. Catenacci and Giuppomi (2009) review the use of Bayesian networks to model and express uncertainty in climate change to aid policy development. Peter et al. (2009) develop a Bayesian network that links the impacts of projected climate change in southern Africa to irrigated agriculture, water storage planning, and biofuel production. All of the above belong to the first category of learning. Furthermore, Bayesian networks are used in these cases to *represent and use* known causal relationships rather than to *discover* causal relationships. Work in the *second category* learns the structure of the Bayesian networks from data using score-based learning algorithms for the purpose of forecasting purposes and do not focus on discovering causal relationships. The works of Cofino et al. (2002), Cano et al. (2004), and Lee and Joseph (2006) fall into this category. All three of them develop Bayesian networks for precipitation forecasting and all of them use modifications of the K2 algorithm, a score-based structure learning algorithm. The *third and final category* uses causal discovery methods for structure learning. For example, Chu et al. (2005) apply structure learning to find the causal structure among time series of remote geospatial indices of ocean surface temperatures and pressures. Chu and Glymour (2008) apply similar methods to study the relationships between four ocean climate indices. Both studies focus on extending standard causal discovery algorithms [such as the PC algorithm by Spirtes and Glymour (1991) used here] to develop causal models based on nonlinear time series. Other work in the third category includes Kennett (2000) (see also Kennett et al. 2001), which derives models for sea breeze prediction using some of the same causal discovery algorithms applied in this paper, although the end product of Kennett's research is again a model for prediction, not causal hypotheses. While the work discussed above—with the exception of Chu et al. (2005) and Chu and Glymour (2008)—consider static models, Cossention et al. (2001) develops a temporal (a.k.a. *dynamic*) Bayesian network for air pollution prediction for the city of Palermo, using expert knowledge and trial-and-error to develop the structure.

Since the early 1980s, the amount of meteorological and climate data collected has been growing every year, probably exponentially (Kenward 2011). In addition to traditional meteorological measurements of local pressure, wind, temperature, and humidity, ground- and space-based remote sensing instruments such as Doppler radar and satellites monitor the states of clouds, precipitation, sea ice coverage, aerosol concentrations, and even acres burned by forest fire. The abundance of available data for a great variety of atmospheric, land and oceanic variables makes it feasible to discover causal relationships from these data. And there is thus great potential in the future for causal discovery to yield new insights for problems of interest to the climate science community. In this paper, we seek to provide an accessible introduction to the topic of causal discovery for climate scientists. The analysis focuses on standard algorithms for causal discovery, applying to nontemporal as well as time series data. The specific technique to be adopted is the so-called constraint-based structure learning, which typically uses a series of conditional independence (CI) tests to detect independence relationships,

and the results are described in the form of graphs. Following this section, we first provide an introduction to causal reasoning in section 2 and to structure learning in section 3. Section 4 provides a detailed case study that demonstrates the causal discovery process step by step. Section 5 presents conclusions and future work with an emphasis on defining new climate networks through causal discovery.

2. Basics of causal reasoning

This section introduces general concepts of causal reasoning, and section 3 describes how they can be used for constraint-based learning.

a. Probabilistic graphical models

Graphs are a convenient way to represent and *visualize* conditional independencies between random variables. Graphs also represent a convenient computational structure that encodes the dependencies in a compact way for use in a great variety of computational algorithms. A *graph* $G = (V, E)$ consists of a set of vertices V and a set of edges E that connect pairs of vertices. *Directed* graphs have a unique direction assigned to each of the edges, while *undirected* graphs have no direction assigned to any of the edges. A directed graph is *acyclic* if it does not contain cycles, that is, starting at any node and following the arrow directions one can never get back to the start node. The vertices of a graph are often called *nodes*. The set of nodes that share an edge with node X in a graph are called the *neighbors* of X. In an undirected graph one only speaks of neighbors. In a directed graph one distinguishes between child and parent nodes. If X and Y are neighbors in a directed graph and the arrow points from X to Y, then X is called a *parent* of Y and Y is called a *child* of X.

Probabilistic graphical models combine tools from graph theory with probability theory. Such models are popular for systems containing uncertainty. The most common type is the *Bayesian Network*, also known as Bayes Net or Belief Network. A Bayesian Network model consists of a directed acyclic graph (DAG) and a probability distribution assigned to each node that defines the probability of the node's state based on the states of its parents (for more details see Charniak 1991; Jensen and Nielsen 2007; Neapolitan 2003). The *Markov Network*, also known as Markov Random Field, is a probabilistic graphical model based on an undirected graph. A Markov network can represent certain dependencies that a Bayesian network cannot (such as bidirectional and cyclic dependencies); on the other hand, it cannot represent certain dependencies that a Bayesian network can—such as the v structures that will be discussed later (see Koller and Friedman 2009 for more details).

Probabilistic graphical models provide an efficient way to represent joint probabilities, in particular if the links represent causal connections. In fact if a system's joint probability can be properly represented by a Bayesian Network, and the edges in the directed graph are based completely on causal relationships, that Bayesian Network generally provides the most compact way of representing the system's joint probability. [The only exception is if the parameters are degenerate in a certain way, namely if the probability distribution violates the *faithfulness* assumption, see for example Spirtes et al. (2000) or Koller and Friedman (2009) for details.] In other words, the underlying graph generally requires the least number of edges and the associated probability tables generally require the least number of probabilities to define the full model. (Note that the causal model tends to be minimal, but a minimal model is not necessarily causal, since there can be more than one minimal model.)

Within the scope of this paper we do not deal explicitly with any of the probabilities. We care only about the structure of the underlying graphs and thus adopt the structure learning algorithms that were developed for graphical models to learn those graphs. Under the conditions to be discussed in section 3 probabilistic graphical models can be interpreted as causal models. For example, in a Bayesian network the arrows of the directed graph can under those conditions often be interpreted as going from *cause* to *effect*. In a Markov model the edges of the graph are undirected so causal influences may go in both directions. In contrast a *correlation graph*, also called correlation network, does not focus on representing causal pathways. In a correlation graph any two nodes are connected if the cross correlation of the data associated with those two nodes is beyond a threshold cc_{min}. Correlation graphs are often used in climate science, and it is always useful to compare graphs obtained from causal reasoning to correlation graphs.

b. The match example

The following match example illustrates several concepts from causal reasoning. One can light a match by striking its head on sand paper. The friction between sand paper and match head causes heat, which in turn starts a chemical reaction in the match head, setting the match on fire. This process can be described by three variables:

- SPaper (yes/no), which indicates whether the match head recently touched the sand paper;
- Temp (low/high), which indicates the temperature of the match head; and
- Fire (yes/no), which indicates whether the match was set on fire.

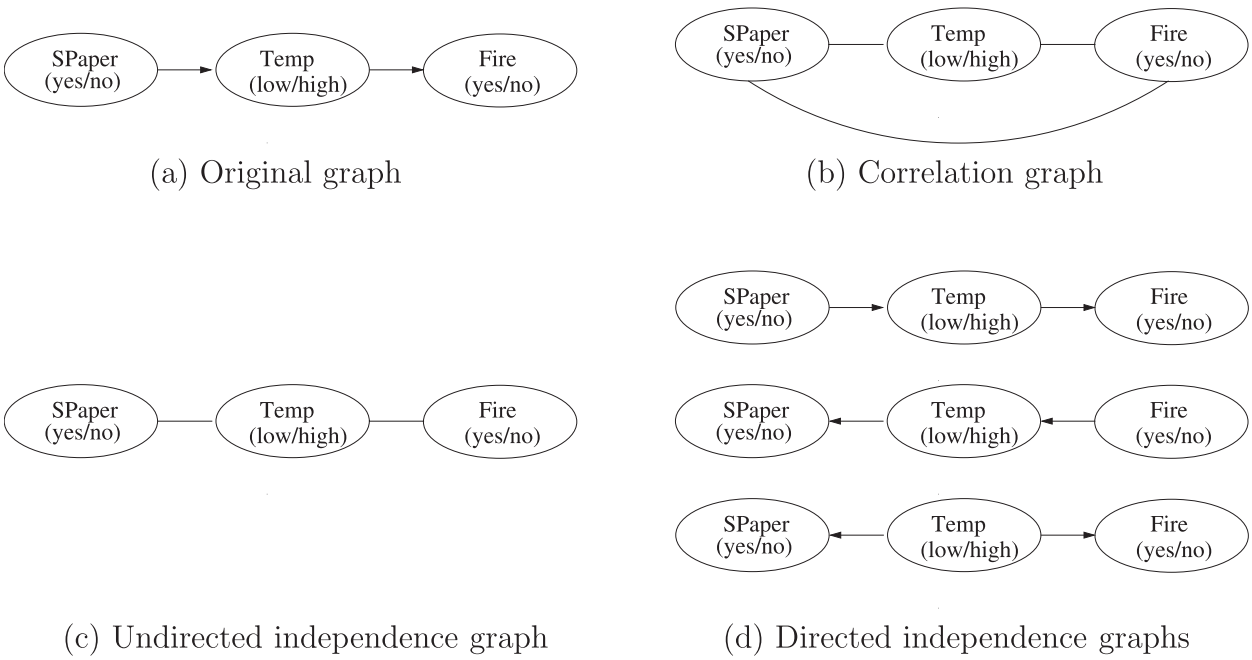


FIG. 1. Match Example.

By understanding the physical processes involved we can describe the causal connections intuitively in the graphical form shown in Fig. 1a. Note that Fig. 1a shows arrows from SPaper to Temp and from Temp to Fire. However, there is *no* edge between SPaper and Fire because the cause–effect relationship between SPaper and Fire *always* goes through the variable Temp. In other words, if we want to make a prediction for whether the match is on fire, and we already know the temperature of the match head, we do not gain any additional information by knowing whether the match recently touched the sand paper. In essence the variable Temp blocks the information flow from SPaper to Fire. In probabilistic terms we say that random variable Fire is *conditionally independent* of SPaper given Temp.

c. Independence and conditional independence

Since it is well known that correlation of two variables does *not* imply causation, tests other than cross correlation must be used to identify potential causal relationships. The basis of causal discovery is to use—in addition to the common independence tests that only involve two variables—also *conditional* independence tests that involve three or more variables.

Two discrete random variables, X and Y, are said to be *independent* of each other if $P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$ for any x, y. Denoting as $P(X = x \mid Y = y)$ the conditional probability that X takes the state x, conditioned on the fact that Y is in state y, two discrete random variables, X and Y, are conditionally independent given a third random variable, Z, if $P(X = x \mid Y = y, Z = z) = P(X = x \mid Z = z)$ for any x, y, and z with $P(Z = z) > 0$. If X and Y are conditionally independent given Z, then if one is interested in the state of X and already knows the state of Z, knowing Y in addition does not add *any* new information. In other words Z blocks the information flow from X to Y. The definition of conditional independence applies not only if Z represents a single random variable, but also for a *set* of several random variables, $Z = \{Z_1, \ldots, Z_k\}$. Although defined here only for discrete variables for the sake of simplicity, the above definitions generalize to continuous variables.

We saw an example of a conditional independence relationship in the match example above (Fire is conditionally independent of SPaper given Temp). In this example the conditional independence was concluded from our understanding of the physical problem. However, in structure learning we want to learn unknown conditional independencies in a system based on data. For that we need tests for independence and CI.

A great variety of measures can be used to test for independence and conditional independence, see Borgelt (2010) for a review. Ideally, any such measure is supposed to yield a value of zero if the variables are (conditionally) independent and nonzero otherwise. In statistics the traditional choice is cross correlation as measure for independence and partial correlation for conditional independence. In theory, partial correlation is an ideal CI

measure only if all variables involved are multivariate Gaussian, but in practice it seems to provide a decent approximation in most cases. Partial correlation has the important advantage that it readily applies to continuous-valued variables, which are quite common in climate research. In information theory the most common choice is mutual information as measure for independence and conditional mutual information for conditional independence. Those measures do not rely on any assumptions on the variables and tend to be a good choice for variables that are discrete by nature. However, they do not readily apply to continuous-valued variables and often do not work well if a variable must be discretized first, especially for coarse discretizations. For this reason partial correlation is used in the case study in section 4, which deals with continuous variables. For a definition of partial correlation, see for example Kachigan (1991).

A special case is as follows: if only time series data is considered and a temporal causal model is desired and no significant preknowledge is available, then the CI test using partial correlation becomes quite similar to the Granger causality test for multivariate time series. (For a discussion of the subtle differences between the concept of Granger causality and Pearls causal model applied for time series data, see White et al. 2011.) In fact one can use the approach by Swanson and Granger (1997) as a short cut to evaluate the CI tests in this case. Their approach is to first calculate a vector autoregression (VAR) model from the data, which describes the current state of all variables in terms of the past evolution of all variables. The coefficients of the VAR model can be used to calculate the partial correlation of each node pair with the linear influence of all other variables removed. The process involves inverting the covariance matrix so care must be taken that it is not close to singular, especially if there are many variables. The partial correlation approximations thus obtained are used directly in the CI tests, instead of calculating partial correlation directly from the data (Swanson and Granger 1997; Eichler 2007).

In practical use CI tests face some additional limitations. Even if two variables are perfectly conditionally independent in theory, because of the noise in the statistical data, CI test results will rarely come out to be exactly zero. Thus *all* CI tests are used in combination with a threshold that determines when variables are considered to be independent. Furthermore, the reliability of the CI test depends on the sample size. The more samples are available the more reliable the result. Finally, for the CI tests calculated without a VAR model, reliability declines rapidly with increasing number k of conditioning variables $Z_1, \ldots, Z_k$, so large conditioning sets should be avoided.

3. Structure learning through CI tests

There are two primary methods for structure learning. The first method is a score-based search that learns the graphs along with probabilities and uses some type of optimization routine to maximize the fit of the model. The most popular algorithm is the K2 algorithm by Cooper and Herskovitz (1992). Numerous other score-based algorithms exist, see Neapolitan (2003). The second method, constraint-based learning, breaks the learning process of a graphical model up into two steps. First CI tests are used to learn as much as possible about the structure of the underlying graph. Once a graph structure is established the probability parameters are learned in the second step. To discover causal hypotheses we only care about the graph structure, so we can simply stop the learning process after the first step and thus never deal with any probability parameters. Both methods have been used successfully for structure learning. We personally prefer the second method because we find its decision-making process more transparent, and we never have to deal with the probabilities. Thus in the remainder of this paper we focus on constraint-based learning as method for structure learning.

We denote the directed and undirected graphs obtained through structure learning as independence graphs because they represent the (conditional) independence relationships. In the four-mode example discussed in section 4 we are most interested in directed graphs, while for other types of climate applications such as climate networks (e.g., Tsonis and Roebber 2004; Tsonis et al. 2006) we may be more interested in undirected graphs. Thus structure learning for both directed and undirected graphs is reviewed here.

a. *Footprints of causal relationships in data*

To recover potential causal relationships from data we need to learn to read their footprints, that is, the traces they leave in the data. There are two main concepts to understand:

(i) the difference between direct and indirect connections and
(ii) so called v structures

Section 3b illustrates the first of these concepts, and section 3c illustrates the second.

b. *Testing for direct connections*

To understand how structure learning with CI tests may work and why, we revisit the match example. For the moment let us forget everything we know about the physical mechanisms in the match example. Instead we are given statistical data obtained by observing the three

variables over an extended amount of time. There is some uncertainty in the system. For example heat may be generated occasionally through other causes, for example, by someone holding the match close to another flame once in a while, or the friction on the sand paper may not be sufficient to start the flame. We now have a large database of observed cases, where each case lists the state of all three random variables. Our task is to learn a graphical model from the data.

First we try the correlation graph. The data would reveal SPaper to be closely correlated with Temp and Temp to be closely correlated with Fire. As a result SPaper is also closely related to Fire, resulting in the correlation graph in Fig. 1b, where all nodes are connected to each other and none of the arrows have a direction associated with them.

Now let us apply CI tests. Since this example only has three nodes only three CI tests need to be performed; namely, we would test whether any two of the variables are conditionally independent given the third variable. For large enough sample size only one CI test would come back negative, namely, only SPaper and Fire are conditionally independent given Temp. This makes intuitive sense because we already know that if we want to know whether the match is likely on fire, and we already know the temperature of the match, it does not matter whether the match recently touched the sand paper. Based on that CI test result we can now eliminate the edge between SPaper and Fire and obtain the undirected independence graph in Fig. 1c.

Learning a directed graph from the CI tests yields the three graphs shown in Fig. 1d. On the top is the correct graph, identical to the one we intuitively came up with in Fig. 1a. The other two vary in the direction of at least one edge. Just based on data it is actually not possible to determine which of the three graphs in Fig. 1d is correct. The three graphs are indistinguishable from a structure learning perspective. One says they are *Markov equivalent*, a concept explained later. Note that the graph with both arrows pointing toward Temp is not included in Fig. 1d. That graph is actually eliminated because the data does not show a v structure, as explained in section 3c.

The match example is very simple, but it demonstrates a basic principle of how CI tests can be used to eliminate one or more edges from a graph.

c. *Finding edge directions through v structures*

A v structure in a directed graph is a child node that has at least two parents that are not connected to each other. In causal reasoning, v structures, also known as *unshielded colliders*, play a key role because they are the key indicators for the *direction* of causal relationships. The following application provides an example of a v structure.

Whether a person develops lung cancer depends among other things on age and smoking habits. In other words the variables Age and Smoking are causes (parents) of the effect (child) LungCancer. Furthermore, let us say that for the considered population the age of a person does not significantly impact whether he/she smokes or not. Thus Age and Smoking are considered independent of each other, and the intuitive causal graph shown on the left of Fig. 2a does not show an edge between them.

The graph in Fig. 2a contains a v structure at LungCancer since this node has two parents that are not connected to each other. The name v structure comes from the fact that these three nodes form the shape of a "V" if we follow the convention of placing parents higher up on the page than children. The v structures leave a distinct footprint that can be detected in the corresponding data and thus can be used to determine directions in a directed graph representation. Namely, the parent nodes are *independent* of each other, but they become *conditionally dependent* if the state of the child is known. Let us illustrate this conditional dependency using the lung cancer example. We made the assumption that Age is independent of Smoking, that is, knowing the age of a person does not tell me anything about his/her smoking habits. However, if we know the status of the variable LungCancer, say that a person has been diagnosed with lung cancer, then the parent nodes become dependent. For example, knowing that a person with lung cancer diagnosis is of a young age raises the probability that the person is smoking because lung cancer patients often have at least one of the two major risk factors, increased age or smoking.

For undirected graphs, v structures also play a special role. An undirected graph is unable to represent the independence relationships of a v structure, resulting in an additional edge between the parents. Figure 2b shows the correlation graph for this example. Figure 2c shows the undirected and Fig. 2d the directed graph that would be obtained through structure learning.

Because of the v structure, learning yields only one directed independence graph in the lung cancer example (Fig. 2d), and this graph perfectly matches the original graph (Fig. 2a). In contrast three different directed independence graphs were obtained for the match example in Fig. 1d because of the *lack* of a v structure in that application.

Because of the v structure, the undirected independence graph contains one more edge than the directed independence graph, namely between the parent nodes Age and Smoking. While in this particular example the correlation graph has the same number of edges as the directed

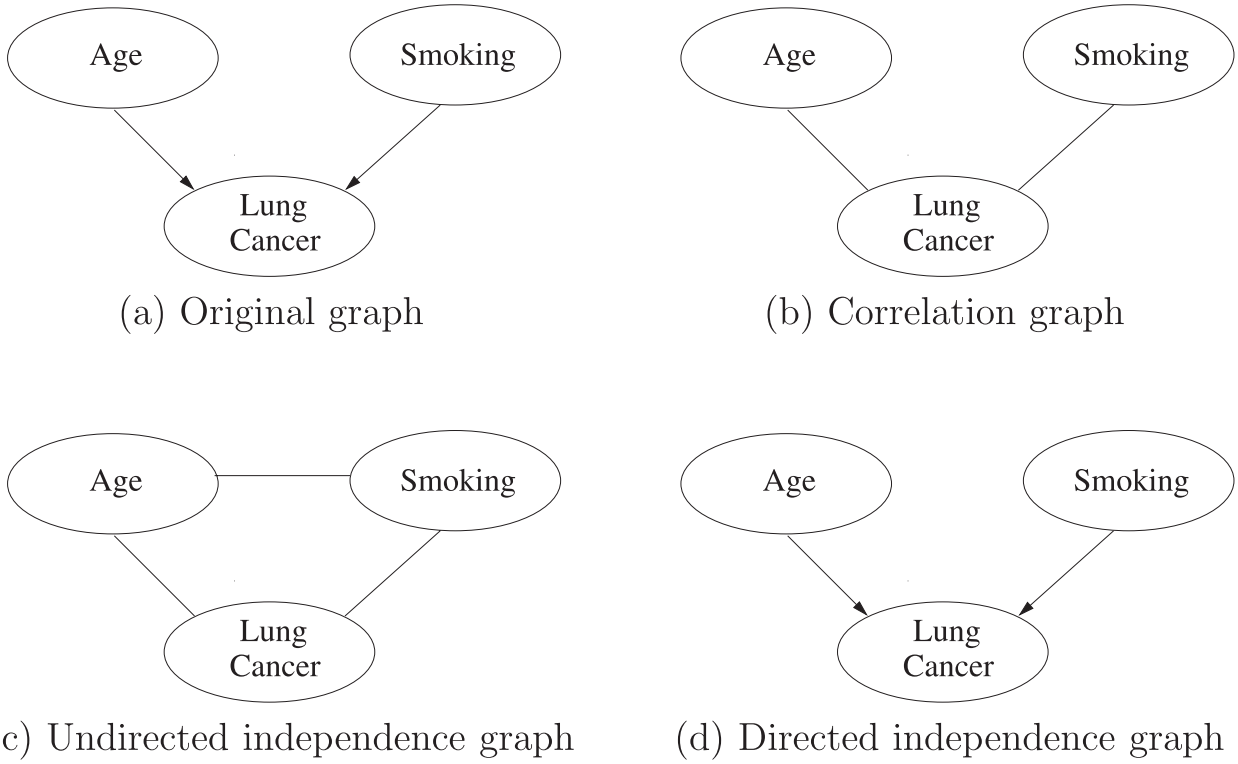


FIG. 2. Lung cancer example.

independence graph, and even one fewer edge than the undirected independence graph, for larger networks the independence graphs generally have significantly fewer edges than the corresponding correlation graphs.

d. PC algorithm

The PC algorithm developed by Spirtes and Glymour (1991) is a basic, but powerful, structure-learning algorithm for directed graphs that is based only on the principles discussed above. It starts out with an undirected graph where any two nodes are connected by an undirected edge. Phase 1 performs edge deletion by considering any pair of nodes, X, Y, and trying to find a set S (not containing X, Y), such that X and Y are conditionally independent *given* S. If such a set can be found, the edge between X and Y is deleted. If no such set can be found the edge remains. Phase 2 adds directions to as many edges as possible first by identifying v structures in the data and adding arrows in the graph accordingly, then by using the constraints that causal loops are not allowed and no additional v structures may be created. The result is a set of Markov equivalent graphs—a concept to be discussed in the next subsection. (If desired the directed graphs can be converted to undirected graphs through a process called *moralization*, see, for example, Koller and Friedman 2009.)

It is quite common in the structure-learning process to merge expert knowledge and automatic-learning algorithms to obtain optimal results. Using as much expert knowledge as possible, such as known direct connections between variables or forbidden edges due to temporal constraints, reduces algorithm complexity and increases the chances of obtaining a valid model. The more complex the model, the more important is it to incorporate any available expert knowledge. Many structure-learning algorithms, including most implementations of the PC algorithm, thus provide the capability of entering preknowledge, such as forced edges or forbidden edges.

e. Markov equivalence and faithfulness of directed graphs

Structure learning from observed data is only able to determine directed graphs up to an equivalence class, namely, the set of Markov equivalent graphs (Verma and Pearl 1990). This equivalence class may contain one or more graphs. Only an intervention analysis—where we actively manipulate the states of some variables in targeted experiments—can reveal additional causal relationships (see Pearl 2000; Murphy 2001).

Two directed graphs are called *Markov equivalent* if they represent the same set of independence relationships. As it turns out, this equivalence can also be

expressed as follows. Two directed graphs are Markov equivalent if they have the same set of edges (ignoring the edge direction) and the same set of v structures. For example, the three directed graphs in Fig. 1d form a Markov equivalence class, and it is not possible to further narrow down which graph is the correct one without performing intervention experiments.

Furthermore, a probability distribution—given by sample data—can be properly represented by a Bayesian network if and only if a DAG exists that is faithful to its probability distribution. A directed acyclic graph is *faithful* to the underlying probability distribution if both represent the same set of conditional independencies. If there is a faithful DAG then the PC algorithm finds it—more precisely its Markov equivalence class. Thus the easiest way to determine whether such a DAG exists is to let the PC algorithm find a model and then check whether it is consistent with the data. Furthermore, if no faithful DAG can be found, the output of the PC algorithms often indicates so by containing (i) undirected edges—indicating there was not enough information to determine the direction; (ii) edges with double arrows—indicating there was contradictory evidence; and (iii) significant inconsistencies for varying sensitivity threshold of the CI tests.

f. From independence graph to causal interpretation

Once we learned independence graphs through structure learning we need to consider under which conditions these graphs can be interpreted in a causal way. There are two types of conditions (for more details see, for example, Koller and Friedman 2009).

1) Going from probability distribution to independence graph, we have to make sure that the obtained independence graph actually models the data well, that is, that it is faithful to the probability distribution.
2) Going from independence graph to causal interpretation, we have to make sure that there are no hidden common causes or other conditions that could cause the independence graph to misrepresent a system's causal relationships.

The first condition roughly translates into the following practical guidelines.

(i) The independence signal must be strong enough to be picked up by the statistical tests in the presence of noise.
(ii) No selection bias is allowed, that is, the data samples must be representative of the independence relationships of the system.
(iii) Probability distributions must be identical and independent. For example, a patient's disease risk for noncontagious diseases are easily modeled, but contagious diseases require work arounds, because one patient's state can affect another patient's state.
(iv) If the independence graph is directed, no causal loops are allowed in the system. If causal loops are present, then a dynamic Bayesian network or a Markov network (undirected graph) should be used.

To meet the second condition, the primary concern is to make sure that the nodes in the graph are causally sufficient, that is, if any two nodes X, Y of the graph have a common cause Z, then Z must also be included in the graph. This condition is sometimes hard to meet in practice because there are often many variables, from ENSO to solar flares, that can have a common influence on variables under consideration. It may be impossible to include them all because of complexity and because some of them cannot even be observed. Algorithms such as the fast causal inference (FCI) algorithm developed by Spirtes and Glymour (1991) can identify the presence of these latent variables under certain conditions but are of high computational complexity and currently not yet feasible for large graphs. Improvements have been suggested, see Colombo et al. (2012), and may help in the future. For now we take the pragmatic approach of using the PC algorithm and interpreting the results accordingly. Namely, we need to consider the possibility that any link detected by the PC algorithm may either present a direct causal connection, be due to a common cause, or a combination of the two. That is why we call the results from the analysis ''causal hypotheses,'' and they must be tested one by one by a domain expert. The contribution of the causal discovery process as described here is therefore to reduce the number of causal hypotheses to a manageable set that can then be tested by a domain expert.

Finally, trends in the data should be removed beforehand. In a way, this is a special case of a hidden common cause because time can be seen as a common cause influencing those variables. The solution is to remove trends from the data in the preprocessing.

4. Case study of four-mode problem

This section demonstrates the use of causal discovery algorithms for an example in climate science. All steps of the causal discovery process are shown in detail. The goal of this case study is to discover hypotheses of causal relationships among four prominent modes of atmospheric low-frequency variability in boreal winter—namely, the Western Pacific Oscillation (WPO), Eastern Pacific Oscillation (EPO), Pacific–North America (PNA) pattern, and North Atlantic Oscillation (NAO). These modes, also known as ''atmospheric teleconnections,'' are characterized by synchronized

low-frequency (longer than typical synoptic time scale of a week) fluctuations in the sea level pressure (SLP) or geopotential height fields at different geographical locations (e.g., Wallace and Gutzler 1981; Barnston and Livezey 1987). Some of these modes, for example, NAO and WPO are largely eddy driven (e.g., Benedict et al. 2004; Franzke et al. 2004; Martius et al. 2007; Rivière and Orlanski 2007; Woollings et al. 2008; Riviére 2010; Deng and Jiang 2011), while others such as PNA are partly eddy driven and partly associated with anomalous tropical convective heating, which is often tied to tropical sea surface temperature (SST) variations (e.g., Franzke et al. 2011).

To improve the skill of extended-range weather forecasting, it is crucial to identify external factors (e.g., tropical SST anomalies) that excite these teleconnections and also to understand dynamical/physical processes that determine their life cycle characteristics (e.g., feedback from synoptic-eddy momentum and heat flux; for an excellent review of this topic, please see Dole 2008). Additionally, it is pointed out by Palmer (1999) that to obtain correct time-mean response to enhanced CO_2 forcing in a climate model, the model should have quasi-stationary regimes (i.e., modes of low-frequency variability) that share structural similarity with those in the real atmosphere. Here we take a different perspective and explore the potential causal relationships among these four modes. These relationships, if confirmed, would serve as basis for formulating hypotheses regarding the dynamics that connect these modes, and these hypotheses can be further tested with general circulation models (GCMs). Specifically, we developed two types of models, static models—involving only the four modes—and temporal models—involving the information of the four modes at different time lags.

a. Data

The data used consists of a time series of daily index value for each of the four modes for the period 1 June 1948–31 May 2011, plus monthly ENSO index (i.e., Niño-3.4 SST) data from 1950 to 2011. The daily index values of the modes are based upon centers-of-action in 500-mb geopotential height (http://www.esrl.noaa.gov/psd/forecasts/teleconn/) and calculated using the 500-mb geopotential height of the National Centers for Environmental Prediction (NCEP)–National Center for Atmospheric Research (NCAR) reanalysis (Kalnay et al. 1996; Kistler et al. 2001). The analysis was focused on the December–February (DJF) period. We performed two types of analysis: a static analysis based on monthly values and including ENSO, and a temporal analysis based on daily values and thus excluding ENSO. For the static model we use monthly averages of the daily data for the four modes, plus monthly ENSO index values. For the temporal analysis we use daily data from all years using a sliding window as follows. To generate a zero lag signal we cut out the DJF values from each year and splice the results. To generate a lagged signal with a delay of N days, we move the cutout window on the original data by N days (N can be positive or negative) and splice the results. Thus lagged signals actually use a few days outside of the DJF period.

b. Algorithm

We use the PC algorithm from section 3 implemented in the TETRAD software package (version 4.3.10-3, available at http://www.phil.cmu.edu/projects/tetrad/). TETRAD provides a convenient graphical user interface and a simple way to enter preknowledge. Out of the choices for conditional independence tests available in TETRAD we chose to use Fisher's Z test, which is a statistical test based on partial correlation, thus it works well for continuous variables, especially if they are normally distributed. (The four modes and ENSO are all nearly normally distributed.) There is one free variable α, which indicates the significance level for the conditional independence tests. The default value of α used in most applications is $\alpha = 0.05$ for small sample size. Lower values of α reduce the number of edges, and higher values increase the number of edges in the graph.

c. Static model

For the static model we added as expert knowledge only the fact that ENSO can be a cause of the other four nodes, but not vice versa. This is added as a constraint for the PC algorithm. (Without this preknowledge the graphs were inconsistent for increasing sensitivity.) The results obtained are shown in Fig. 3. For very small values of α (Fig. 3a) only the strongest two links appear, ENSO $\rightarrow$ PNA and WPO—EPO , with the direction of the latter undetermined. For slightly larger values of α (Fig. 3b) a third link appears from NAO to PNA. Increasing α further (Fig. 3c) we get one additional arrow from ENSO to WPO. This edge also causes the edge between EPO and WPO to get a direction (toward WPO) since the algorithm has discovered a v structure at WPO. For even larger values of α (Fig. 3d) an additional arrow from EPO to NAO appears. Simultaneously, the edge between PNA and NAO is now classified as bidirectional since the algorithm identified a v structure at both PNA and NAO.

The static plots are fairly consistent for increasing α, namely, any link appearing for low values of α also

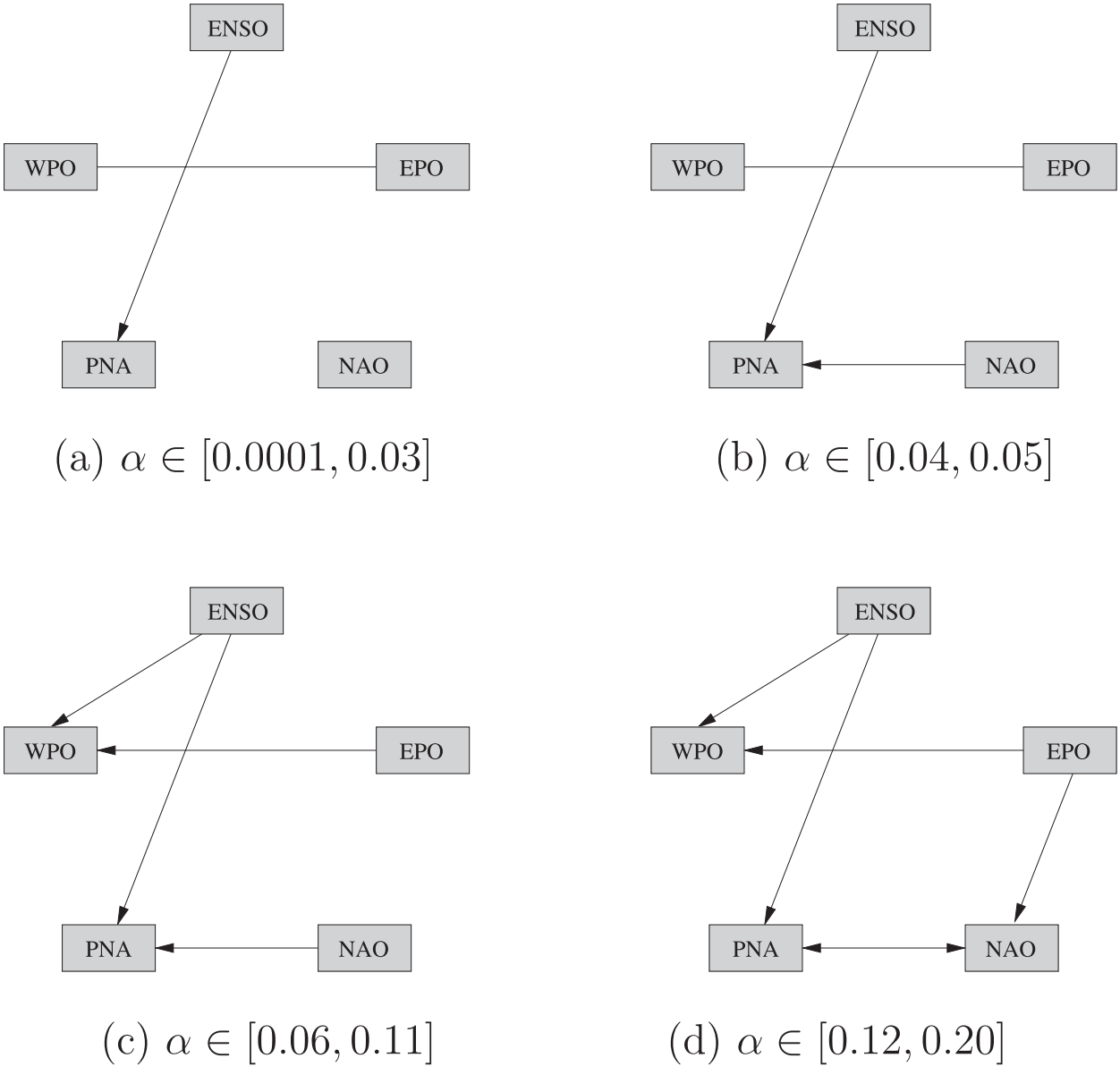


FIG. 3. Static independence graphs from PC algorithm for varying significance level α.

exists for higher values of α. However, one of the links gains a direction (EPO to WPO) for increasing α and another link becomes bidirectional (PNA $\leftrightarrow$ NAO). The fact that preknowledge was required to get consistent results—indicating that the independence signals are very weak—and the fact that one edge is bidirectional—indicating that there may be information flow in both directions—motivated us to consider a temporal model as well. A temporal model can pick up dependencies that are strong on a shorter time scale (less than one month), and it can represent bidirectional dependencies. Both models together can provide a more complete picture of the physical mechanisms in the system. Thus the results for the static model are discussed further once results for the temporal model are obtained in section 4d.

d. Temporal model

In the temporal model the nodes of the graph consist of the daily index values of the four modes at different time lags. Each node is indicated by the mode name followed by the number of lag days. For example, EPO with $N = 3$ days lag is denoted as EPO3 and WPO with $N = -6$ days lag is denoted as WPO-6. However, since ENSO index based upon SST data is only available as

TABLE 1. Sample tier assignment used in TETRAD.

Tier				
1	EPO-21	NAO-21	PNA-21	WPO-21
⋮	⋮	⋮	⋮	⋮
7	EPO-3	NAO-3	PNA-3	WPO-3
8	EPO0	NAO0	PNA0	WPO0
9	EPO3	NAO3	PNA3	WPO3
⋮	⋮	⋮	⋮	⋮
15	EPO21	NAO21	PNA21	WPO21

monthly averages, we had to exclude it from the temporal model.

1) INCORPORATING TEMPORAL CONSTRAINTS

As mentioned before, it is good practice to add available expert knowledge to the structure learning in the form of forced or forbidden edges. For this application we chose to only use temporal constraints, as follows. Since an event in the future cannot have an effect on an event in the past, any node with time index N should only be pointing toward nodes with time index N' of the same time slice or later ($N' \geq N$). We use TETRAD's *tier feature* to encode these edge constraints. We divided the variables into tiers according to their time slices. For example, if we use 15 slices with a distance of 3 days, the tiers are as shown in Table 1. Causal directions are not allowed to go from a lower tier to a higher tier in TETRAD. That means that there can be arrows for example from EPO3 to PNA3, PNA6, PNA9, . . . , but not to PNA0, PNA-3, . . . , which is exactly the constraint we wanted to achieve.

2) CONCEPTS AND IMPORTANT PARAMETERS

Let us denote as *intramode connection* the connection between two nodes that represent different time lags of the same mode, for example, EPO3 → EPO6. In contrast an *intermode connection* denotes the connection between two different modes, for example, EPO3 → WPO6.

A few important parameters remain to be chosen to obtain a temporal graph. The following outlines how to choose them.

(i) α: Threshold for CI tests

One should always run the simulations for various values of α to observe any trends for increasing sensitivity. Most importantly, the model is only trustworthy if arrows present in the graphs of low α are also present in graphs with higher values of α.

(ii) D: Distance between time slices in days

The distance between time slices is very important, since the models tend to pick up only the most important connections. Clearly any mode will have strong connections to itself with a delay of 1, 2, 3, . . . days depending on the persistency of the mode. If D is chosen very small, a large number of intramode connections are included and we may *only* pick up those intramode connections. D can be chosen by expert knowledge (for example if a typical time delay between variables is known) or simply by trial and error.

(iii) S: Number of slices to include in the model

Higher values of S increase algorithm complexity, but in our application that did not seem to be a limiting factor. All models were calculated in minutes.

(iv) S_{del}: Number of time slices to delete at top of graph

In many cases the model needs a few time slices to converge to a proper independence model. The reason is an initialization problem; namely, to determine the causal flow *originating* in a time slice, it is crucial to have information of the causal flow *into* that time slice. Since the first few time slices are lacking that information (because no prior time slices are included), they often yield erroneous links. This problem is easily solved by developing the model for more slices than needed and then deleting the first few slices in the results. How many slices should be deleted is usually obvious from the resulting graph because the first (erroneous) slices usually differ significantly from the stable pattern emerging in the later slices. In our examples it was sufficient to delete the first three slices.

3) RESULTS FOR TEMPORAL MODEL

We performed the analysis for $D = 1$, 2 and 3 days between time slices and for sensitivity values of $\alpha = 0.001$, 0.01 and 0.05. For each combination we calculated an independence graph for lag times ranging from −15 to 15 days (−16 to 16 days for $D = 2$). Figure 4 demonstrates that process for $D = 3$ and $\alpha = 0.01$. First an independence graph is calculated including slices from −24 to 15 days. The result is shown in Fig. 4a. Initialization problems tend to create a large number of additional erroneous links, especially in the very first time slice of the graph. That is exactly the case here. While the number of intermode edges originating in a time slice ranges between 4 and 8 edges for all later slices, the very first slice in Fig. 4a contains 22 such edges, that is, at least 14 of those edges are likely to be erroneous. Eliminating the first three time slices in Fig. 4a results in Fig. 4b. The graph in Fig. 4b still shows irregularities, but those represent varying patterns in the data—either because of noise or because of nonstationary behavior—and do not stem from an initialization problem. In the following we denote

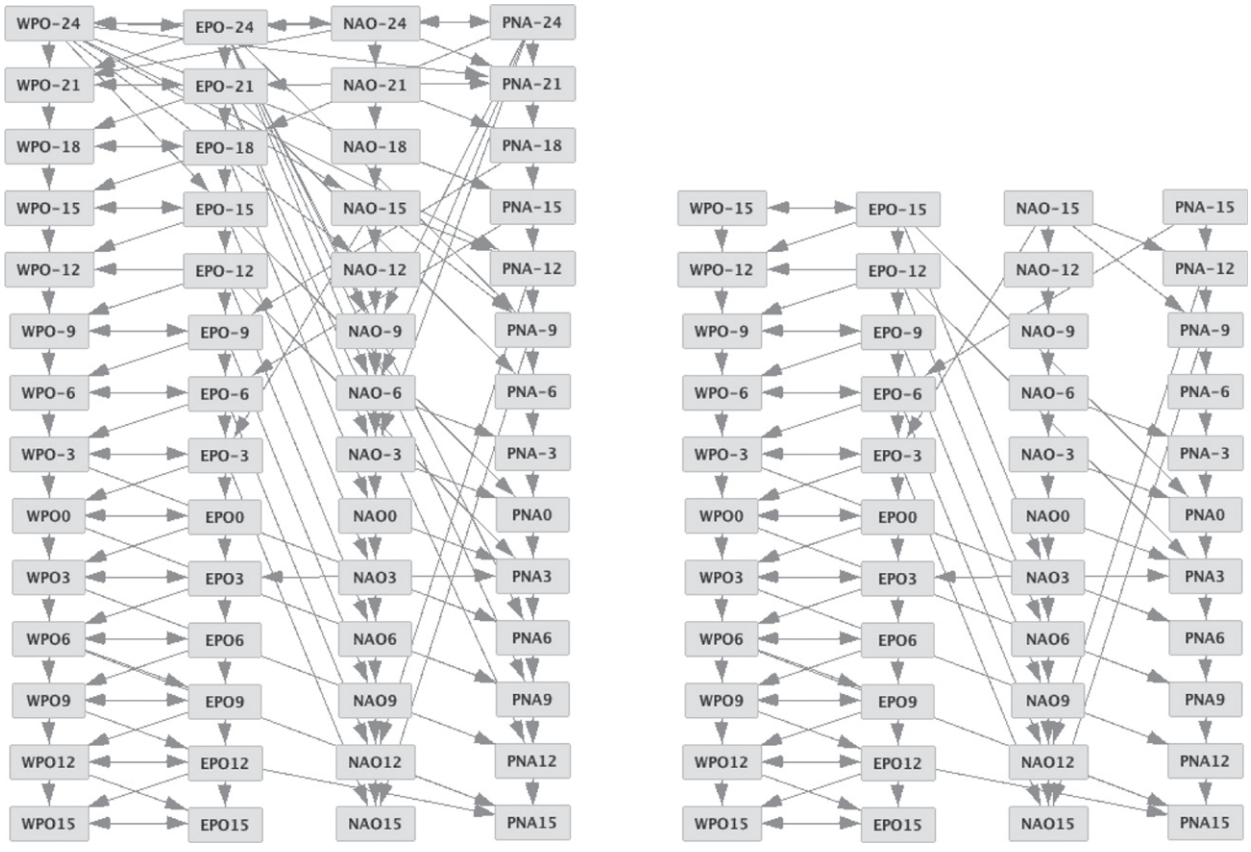


FIG. 4. Temporal independence graph for $D = 3$ days between slices and $\alpha = 0.01$.

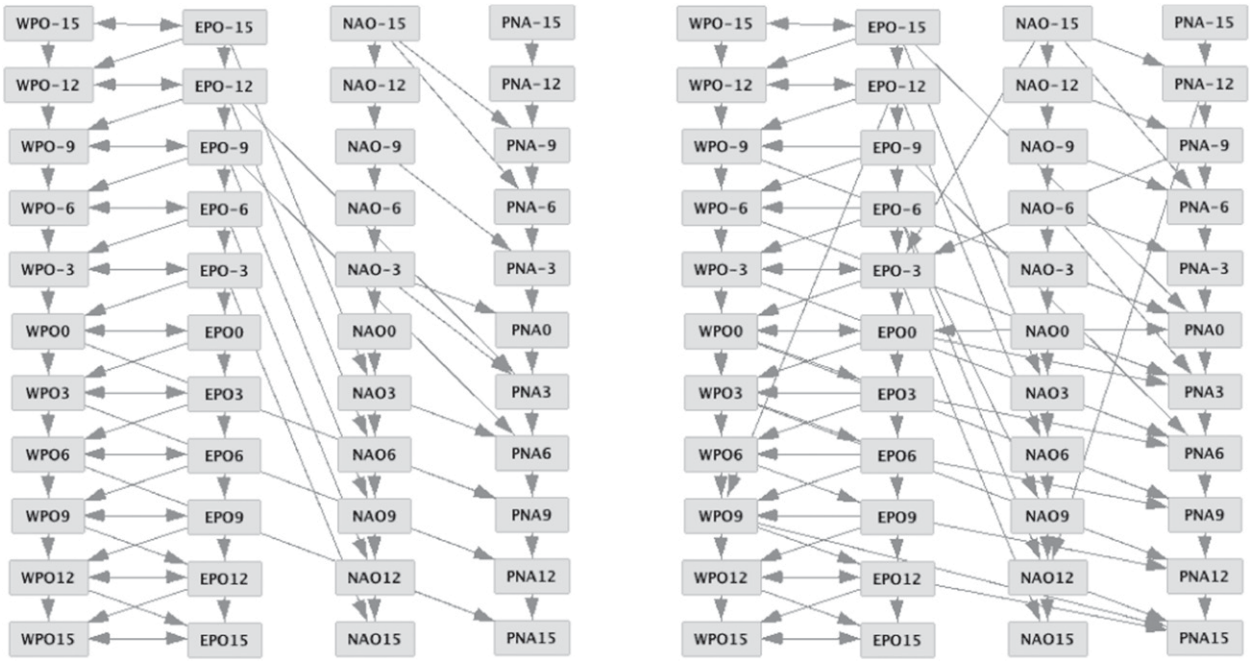
as *stable pattern* the graph with the first three slices removed, even if it is not perfectly regular.

While all intermode connections are visible in Fig. 4b, some of the intramode connections are hidden behind other arrows and node boxes. An analysis of the hidden edges in Fig. 4b shows that each mode tends to be connected to itself one slice and two slices later, for example, EPO3 connects to EPO6 and EPO9. That means that each mode has a persistency (local memory) of about 6 days. A summary of the strongest connections for $D = 3$ is provided (in Fig. 7 below).

Figure 5 shows the stable graphs for $D = 3$ for other values of α. For $\alpha = 0.001$ the result is almost identical to the one for $\alpha = 0.01$ (Fig. 4b). For $\alpha = 0.05$ additional edges appear. Overall the graphs are very consistent for increasing α, that is, connections that appear for smaller α tend to be present also for larger α. The number of edges does not change drastically either—a total of 121, 130, and 142 edges, respectively, for $\alpha = 0.001, 0.01, 0.05$ (counted *after* the first 3 slices are deleted). The persistency (local memory) of each mode also remains constant at 6 days for each node for all three α values.

Figure 6 shows the stable graphs for $D = 1$ days between slices after the first three slices have been deleted. For $D = 1$ the number of edges is much larger, namely, 661 (740) edges for $\alpha = 0.01$ ($\alpha = 0.05$), most of which are intramode edges. The reason is that for example WPO connects to itself with a delay of 3 and 6 days in the graph for $D = 3$, but it connects to itself with a delay of 1, 2, 3, 4, and 5 days for $D = 1$, that is, more than twice as many edges are needed. The many intramode edges tend to dominate the graph for low α, so to detect intermode edges we need to use a larger α value. For example for $\alpha = 0.01$ (Fig. 6a) the only intermode connections detected are between EPO and WPO. For $\alpha = 0.05$ other connections are detected as well. Thus we use in this application as default value $\alpha = 0.01$ for $D = 2, 3$ and $\alpha = 0.05$ for $D = 1$. Note that the graphs in Fig. 6 indicate that the relationship between EPO and WPO changes from a simultaneous connection in the first half of the time slices, to a connection with 1-day delay in the second half.

Figure 7 summarizes the strongest links found for $D = 3$. The strongest links are indicated by a solid line, medium strength links are indicated by a dashed line, and



(a) $\alpha = 0.001$ (b) $\alpha = 0.05$

FIG. 5. Graph for $D = 3$ days between slices and varying α.

weak links are not shown at all. The time delay in number of days is given next to each arrow. For connections with more than one delay value we use the following notation. Listing several time delays separated by a comma implies that there are *multiple edges*, namely *one edge for each time delay listed*. Listing time delays separated by "or" means that there is *only one edge* and that its delay varies between different values—with the dominant values listed first.

The strongest intermode connections present for $D = 3$ are as follows:

(T1) EPO ↔ WPO with a delay of 0 days, where the link EPO → WPO seems to be stronger than the reverse;
(T2) EPO → WPO with a delay of 3 days;
(T3) EPO → NAO with 18 days delay; and
(T4) NAO → PNA with delay ranging from 3 to 6 days.

The strongest intermode connections obtained for $D = 1, 2$ are very similar to (T1)–(T4). The time delays vary a little and some of the medium strength edges differ, but overall there is good agreement with the results above, especially for $D = 1$.

In comparison the static graphs in Fig. 3 indicate the following intermode connections:

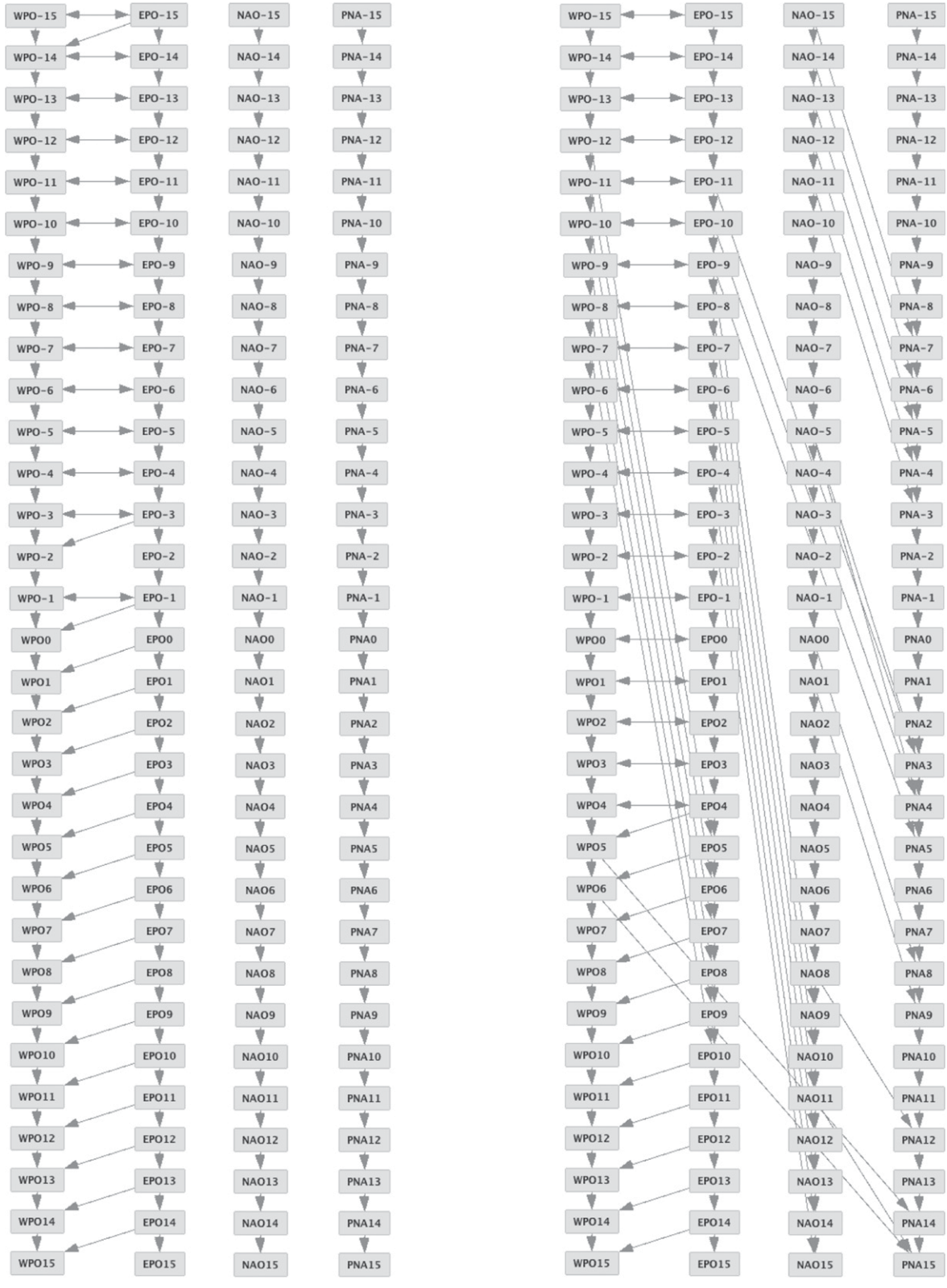
(S1) EPO → WPO (EPO − WPO for some α);
(S2) NAO → PNA (NAO − PNA for some α);
(S3) ENSO is a common cause of WPO and PNA, which should show up as some link WPO − PNA in the temporal graphs without ENSO; and
(S4) EPO → NAO.

The static model is in good agreement with the temporal model since all strong links from the temporal model show up in the static model and vice versa. (An exception is the strong link WPO–PNA from the static model shows up only as a medium link in Fig. 7.) The advantage of the temporal model is that it provides specific time delays and that it is better suited to capture dependencies on a shorter time scale. Additionally, the static model indicates that ENSO appears to be a common factor for only WPO and PNA—all other links in the temporal graphs are unlikely to be due to ENSO being a common cause.

e. Interpretation of results

The temporal independence graphs generated several hypotheses for potential causal pathways, (T1)–(T4). As we know from section 3 for each hypothesis we need to test whether it represents a direct connection, is due to a common cause, or a combination of the two. Nevertheless we have thus narrowed down the number of causal hypotheses to just a few with specific time delays.

After obtaining the above list of specific causal hypotheses, [(T1)–(T4)], connecting pairs of modes, we seek to



(a) $\alpha = 0.01$ (b) $\alpha = 0.05$

FIG. 6. Graph for $D = 1$ day between slices and varying α, first 3 slices already deleted.

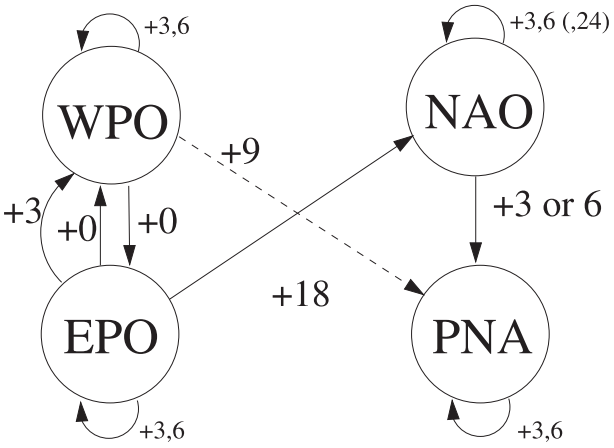


FIG. 7. Summary graph for $D = 3$ and $\alpha = 0.01$. Strong (medium) strength connections are shown as solid (dashed) arrows with corresponding time delays.

identify underlying dynamical mechanisms that would explain those connections and thus support the hypotheses.

For example, the following chain of events can be envisioned as a plausible explanation for the WPO → EPO connection: 1) phase transition in WPO (induced either by anomalous tropical SST forcing or high-latitude blocking, e.g., Woollings et al. 2008; Dole 2008) is closely coupled to changes in the intensity/location of the subtropical jet; 2) variability in the subtropical jet leads to changes in the property (track, strength, etc.) of synoptic eddies of the Pacific storm track that is located downstream of the jet (e.g., Deng and Mak 2005, 2006); 3) anomalous eddy forcing (in terms of vorticity and/or heat flux) drives the geopotential height tendency characteristic of phase transition in EPO. On the other hand, the even stronger EPO → WPO connection with a 3-day delay could be reflecting the fact that WPO is largely eddydriven with forcing mostly originating in the central-eastern Pacific where synoptic eddies attain their maximum intensity, break, and trigger first a phase transition in EPO. The above hypothesis regarding the WPO–EPO connection can be readily tested through controlled experiments with an idealized atmospheric GCM (e.g., Mak and Deng 2006). The EPO to NAO connection, as seen in both the static and temporal model, might be a demonstration of the role of transient eddy forcing (especially over the North American continent) in bridging the variability of two eddy-driven modes over the North Pacific and the North Atlantic (e.g., Li and Lau 2012). The last of the strong links identified in the temporal model, NAO to PNA with a delay of 3 to 6 days, is a new discovery. Previous studies focusing on dynamical processes linking ENSO variability and strength of stratospheric polar vortex have hinted a connection between the two but with an opposite direction, that is, PNA → NAO (e.g., Garfinkel and Hartmann 2008; Hegyi and Deng 2011). The link found here through causal discovery methods, including the time lag, however, is consistent with the result from a recent and independent study that utilized rotated empirical orthogonal function (REOF) analysis (Baxter and Nigam 2012). Whether this connection reflects a downstream, circum-hemispheric modulation of NAO variability on PNA remains to be investigated with a dynamical model.

f. Comparison to correlation graphs

As correlation graphs are much more common in climate science—they are the standard model for *climate network*—it is a legitimate question whether similar information could have been obtained for this application using a correlation graph. Thus we constructed correlation graphs corresponding to the temporal independence graphs for $D = 3$ and time slices -15 to 15. We use the same nodes, but any pair of nodes in the correlation graph is connected by an edge if their correlation exceeds a chosen threshold cc_{min}. Furthermore, we added edge directions in the correlation graph by using the same temporal constraints as in the independence graph. Namely, for any connected node pair of different time slices the direction of the edge is always from the earlier time slice to the later time slice. If nodes lie in the same time slice then no direction is assigned. We denote the result a *temporal* correlation network or graph.

We calculated temporal correlation graphs for $cc_{min} = 0.75, 0.50, 0.25, 0.20, 0.15, 0.1$. For $cc_{min} = 0.75$ there are only intramode connections. For $cc_{min} = 0.50, 0.25, 0.20$, and 0.15 the only intermode connections found are between EPO and WPO with varying time delays. The number of edges increases rapidly for decreasing cc_{min}, so that the graph for $cc_{min} = 0.15$, for example, includes all of the following edges: EPO → WPO $+0, 3, \ldots, 21$ and WPO → WPO $+3, 6, \ldots, 27$. Finally, for $cc_{min} = 0.1$ the following intermode connections appear: EPO—WPO +0; EPO → WPO $+3, 6, \ldots, 30$; WPO → EPO $+3, 6, 12, \ldots, 21$; EPO → NAO +18, and NAO → WPO +9.

Note that the independence graph for $D = 3$ and $\alpha = 0.001$ already indicated all of the major intermode hypotheses, (T1)–(T4), obtained above, and only has an average of 11 edges per time slice. In contrast the correlation graph for $cc_{min} = 0.15$ contains twice as many edges and still does not detect any intermode connection other than those linking EPO and WPO. Here, cc_{min} has to be decreased to a *very low* value, $cc_{min} = 0.1$, to obtain any additional intermode connections, and by then the average number of edges per time slice has grown to 30 (almost three times that of the independence graph) and

many time delays are too broad to yield specific hypotheses. Furthermore, the links found for $cc_{min} = 0.1$ and $D = 3$ do not match the ones from the independence graphs, so assuming that the independence graphs yielded correct results, the correlation graph does not yield any useful information. In summary, the capability of distinguishing between direct and indirect connections—which correlation graphs by nature do not possess—seems crucial to detecting potential causal relationships.

5. Conclusions and future work

Causal reasoning shows tremendous potential to generate new causal hypotheses for problems in climate science. Application of a specific causal discovery algorithm (constraint-based structure learning) to the daily index values of the four prominent teleconnection patterns in the atmosphere yields important new information regarding their potential causal relationship. Among them are the robust simultaneous coupling between WPO and EPO (phase transition in EPO also tends to lead that in WPO by 3 days), EPO to NAO (18-day delay), and NAO to PNA (3–6-day delay). The fact that WPO and EPO are nearly indistinguishable from the cause–effect perspective can be explained through invoking a chain of hypothetical events that involve forced variability in the subtropical jet, storm-track dynamics, and synoptic-eddy feedback to low-frequency flow. Part of our future work will be testing this hypothesis with an idealized atmospheric GCM. The EPO to NAO connection is likely established by anomalous transient eddy forcing that could extend from the eastern North Pacific to North American continent and North Atlantic. The NAO to PNA connection, on the other hand, might indicate a downstream, circum-hemispheric modulation of low-frequency variability over the eastern North Pacific by NAO. The relative importance of synoptic versus low-frequency eddies in this modulation needs further investigation. Compared to correlation graphs that are used more often in climate science, the independence graphs derived here provide a more compact representation of the potential causal relationships within the considered system, with much less ambiguity.

An important extension of our analysis, as part of our future work, is to introduce independence graphs to the area of climate networks, an area that currently uses primarily correlation graphs. In their seminal papers Tsonis and Roebber (2004) and Tsonis et al. (2006) introduced the idea of *climate networks*, which brought tools from network analysis to the field of climate science. Their basic idea is to use atmospheric fields—or other physical quantities—to define a correlation network of nodes, where each node represents a point on a global grid. Any two nodes are connected if the cross correlation of the data associated with those two nodes is beyond a threshold cc_{min}. Since these correlation networks were introduced to climate science in 2004, there has been a flurry of research activity in this area, discussing definition, calculation, evaluation, and interpretation of climate networks (Tsonis et al. 2006; Donges et al. 2009). Several research groups related global network changes over a longtime scale to El Niño activity (Tsonis et al. 2007; Tsonis and Swanson 2008; Gozolchiani et al. 2008; Yamasaki et al. 2008, 2009). A summary of the progress, opportunities, and challenges of networks in climate science was presented by Steinhaeuser et al. (2010). While most climate networks are defined as correlation networks, two other definitions have recently been proposed, mutual information (MI) networks (Donges et al. 2009) and phase synchronization networks (Yamasaki et al. 2009). All three network definitions, however, decide whether an edge exists between two nodes in the network based only on a test involving those two nodes and the results are fairly similar for all three. We believe that using independence graphs based on structure learning for climate networks would yield networks with significantly fewer edges by eliminating indirect connections. Since edge *directions* are hard to determine in such a large network, undirected graphs (Markov networks) are likely to be the best choice for climate networks. Furthermore, while correlation-based climate networks focus on *similarity* between nodes, independence graphs would provide an alternative viewpoint by focusing on *information flow* within the network.

Acknowledgments. We thank four anonymous reviewers for their thoughtful comments and suggestions that led to major improvement of the manuscript. The NCEP–NCAR reanalysis data used in this study was provided through the NOAA Climate Diagnostics Center. This research was in part supported by the DOE Office of Science Regional and Global Climate Modeling (RGCM) program under Grant DE-SC0005596 and NASA Energy and Water Cycle Study (NEWS) program under Grant NNX09AJ36G.

REFERENCES

Abramson, B., J. Brown, W. Edwards, M. Murphy, and R. Winkler, 1996: Hailfinder: A Bayesian system for forecasting severe weather. *Int. J. Forecasting,* **12,** 57–71.

Arnold, A., Y. Liu, and N. Abe, 2007: Temporal causal modeling with graphical Granger methods. *Proc. 13th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD '07),* San Jose, CA, ICKDD, 10 pp. [Available online at http://www.cs.cmu.edu/~aarnold/cald/frp781-arnold.pdf.]

Barnston, A. G., and R. E. Livezey, 1987: Classification, seasonality, and persistence of low-frequency atmospheric circulation patterns. *Mon. Wea. Rev.,* **115,** 1083–1126.

Baxter, S., and S. Nigam, 2012: Pentad analysis of wintertime PNA development and its relationship to the NAO. Preprints, *24th Conf. on Climate Variability and Change,* New Orleans, LA, Amer. Meteor. Soc., P202152. [Available online at http://ams.confex.com/ams/92Annual/webprogram/Paper202152.html.]

Benedict, J., S. Lee, and S. Feldstein, 2004: A synoptic view of the North Atlantic Oscillation. *J. Atmos. Sci.,* **61,** 121–144.

Borgelt, C., 2010: A conditional independence algorithm for learning undirected graphical models. *J. Comput. Syst. Sci.,* **76,** 21–33.

Cano, R., C. Sordo, and J. Gutierrez, 2004: Applications of bayesian networks in meteorology. *Advances in Bayesian Networks,* J. A. Gámez et al., Eds., Springer, 309–327.

Catenacci, M., and C. Giuppomi, 2009: Potentials of bayesian networks to deal with uncertainty in climate change adaptation policies. Centro Euro-Mediterraneo per i Cambiamenti Climatici (CMCC) Tech. Rep. RP0070, 29 pp.

Charniak, E., 1991: Bayesian networks without tears. *AI Mag.,* **12** (4), 50–63.

Chu, T., and C. Glymour, 2008: Search for additive nonlinear time series causal models. *J. Mach. Learn. Res.,* **9,** 967–991.

——, D. Danks, and C. Glymour, 2005: Data driven methods for nonlinear granger causality: Climate teleconnection mechanisms. Carnegie Mellon University, Dept. of Philosophy Tech. Rep. CMU-PHIL, 171 pp.

Cofino, A., R. Cano, C. Sordo, and J. Gutierrez, 2002: Bayesian networks for probabilistic weather prediction. *Proc. 15th European Conf. on Artifical Intelligence (ECAI 2002),* Lyon, France, ECAI, 695–700.

Colombo, D., M. H. Maathuis, M. Kalisch, and T. S. Richardson, 2012: Learning high-dimensional directed acyclic graphs with latent and selection variables. *Ann. Stat.,* **40,** 294–321.

Cooper, G., and E. Herskovitz, 1992: A Bayesian method for the induction of probabilistic networks from data. *Mach. Learn.,* **9,** 330–347.

Cossention, M., F. Raimondi, and M. Vitale, 2001: Bayesian models of the pm 10 atmospheric urban pollution. *Proc. Ninth Int. Conf. on Modeling, Monitoring and Management of Air Pollution: Air Pollution IX,* Ancona, Italy, Wessex, 143–152.

Deng, Y., and M. Mak, 2005: An idealized model study relevant to the dynamics of the midwinter minimum of the Pacific storm track. *J. Atmos. Sci.,* **62,** 1209–1225.

——, and ——, 2006: Nature of the differences in the intraseasonal variability of the Pacific and Atlantic storm tracks: A diagnostic study. *J. Atmos. Sci.,* **63,** 2602–2615.

——, and T. Jiang, 2011: Intraseasonal modulation of the North Pacific storm track by tropical convection in boreal winter. *J. Climate,* **24,** 1122–1137.

Dole, R., 2008: Linking weather and climate. *Synoptic-Dynamic Meteorology and Weather Analysis and Forecasting: A Tribute to Fred Sanders,* Amer. Meteor. Soc., 297–348.

Donges, J., Y. Zou, N. Marwan, and J. Kurths, 2009: The backbone of the climate network. *Europhys. Lett.,* **87,** 48007, doi:10.1209/0295-5075/87/48007.

Eichler, M., 2007: Granger causality and path diagrams for multivariate time series. *J. Econom.,* **137,** 334–353.

Franzke, C., S. Lee, and S. Feldstein, 2004: Is the North Atlantic Oscillation a breaking wave? *J. Atmos. Sci.,* **61,** 145–160.

——, S. Feldstein, and S. Lee, 2011: Synoptic analysis of the Pacific–North American teleconnection pattern. *Quart. J. Roy. Meteor. Soc.,* **137,** 329–346.

Friedman, N., M. Linial, I. Nachman, and D. Pe'er, 2000: Using bayesian networks to analyze expression data. *J. Comput. Biol.,* **7** (3–4), 601–620.

Garfinkel, C., and D. Hartmann, 2008: Different ENSO teleconnections and their effects of the stratospheric polar vortex. *J. Geophys. Res.,* **113,** D18114, doi:10.1029/2008JD009920.

Gozolchiani, A., K. Yamasako, O. Gazit, and S. Havlin, 2008: Pattern of climate network blinking links follws El Niño events. *Europhys. Lett.,* **83,** 28005, doi:10.1209/0295-5075/83/28005.

Granger, C. W. J., 1969: Investigating causal relations by econometric models and cross-spectral methods. *Econometrica,* **37,** 424–438.

Hegyi, B. M., and Y. Deng, 2011: A dynamical fingerprint of tropical Pacific sea surface temperatures on the decadal-scale variability of cool-season arctic precipitation. *J. Geophys. Res.,* **116,** D20121, doi:10.1029/2011JD016001.

Jensen, F. V., and T. D. Nielsen, 2007: *Bayesian Networks and Decision Graphs*. 2nd ed. Springer, 284 pp.

Kachigan, S., 1991: *Multivariate Statistical Analysis*. 3rd ed. Radius Press, 303 pp.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.,* **77,** 437–471.

Kennett, R. J., 2000: Seabreeze prediction using bayesian networks. Honours thesis, School of Computer Science and Software Engineering, Monash University, 48 pp. [Available online at http://www.csse.monash.edu.au/hons/projects/2000/Russell.Kennett/thesis.ps.]

——, K. B. Korb, and A. E. Nicholson, 2001: Seabreeze prediction using bayesian networks. *Proc. Fifth Pacific-Asia Conference on Knowledge Discovery and Data Minung (PAKDD'01),* Hong Kong, China, PAKDD, 148–153.

Kenward, A., cited 2011: Data storm: What to do with all this climate information? [Available online at http://www.climatecentral.org/blogs/data-storm-what-to-do-with-all-this-climate-information/.]

Kistler, R., and Coauthors, 2001: The NCEP–NCAR 50-Year Reanalysis: Monthly means CD-ROM and documentation. *Bull. Amer. Meteor. Soc.,* **82,** 247–267.

Koller, D., and N. Friedman, 2009: *Probabilistic Graphical Models–Principles and Techniques*. 1st ed. MIT Press, 1280 pp.

Lee, B., and J. Joseph, 2006: Learning a probabilistic model of rainfall using graphical models, project report for machine learning (Fall 2006). Carnegie Mellon University School of Computer Science, 8 pp. [Available online at http://www.cs.cmu.edu/~epxing/Class/10701-06f/project-reports/lee_joseph.pdf.]

Li, Y., and N.-C. Lau, 2012: Impact of ENSO on the atmospheric variability over the North Atlantic in late winter–role of the transient eddies. *J. Climate,* **25,** 320–342.

Mak, M., and Y. Deng, 2006: Diagnostic and dynamical analyses of two outstanding aspects of storm tracks. *Dyn. Atmos. Oceans,* **43** (1–2), 80–99, doi:10.1016/j.dynatmoce.2006.06.004.

Margolin, A. A., I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano, 2006: Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinf.,* **7** (Suppl.), S7, doi:10.1186/1471-2105-7-S1-S7.

Martius, O., C. Schwierz, and H. Davies, 2007: Breaking waves at the tropopause in the wintertime Northern Hemisphere: Climatological analyses of the orientation and the theoretical lc1/2 classification. *J. Atmos. Sci.,* **64,** 2576–2592.

Murphy, K. P., 2001: Active learning of causal Bayes net structure. University of California, Berkeley, Department of Computer Science Tech. Rep., 8 pp. [Available online at http://www.cs.ubc.ca/~murphyk/papers/alearn.ps.gz.]

Neapolitan, R. E., 2003: *Learning Bayesian Networks*. Prentice Hall, 647 pp.

Palmer, T., 1999: A nonlinear dynamical perspective on climate prediction. *J. Climate,* **12,** 575–591.

Pearl, J., 1988: *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. 2nd ed. Morgan Kaufman Publishers, 552 pp.

——, 2000: *Causality–Models, Reasoning and Inference*. Cambridge University Press, 400 pp.

Peter, C., W. de Lange, J. Musango, K. April, and A. Potgleter, 2009: Applying Bayesian modelling to assess climate change effects on biofuel production. *Climate Res.,* **40,** 249–260.

Rebane, G., and J. Pearl, 1987: The recovery of causal poly-trees from statistical data. *Proc. Sixth Workshop on Uncertainty in AI,* Seattle, WA, AAAI, 222–228.

Riviére, G., 2010: Role of Rossby wave breaking in the west Pacific teleconnection. *Geophys. Res. Lett.,* **37,** L11802, doi:10.1029/2010GL043309.

——, and I. Orlanski, 2007: Characteristics of the Atlantic storm-track eddy activity and its relation with the North Atlantic oscillation. *J. Atmos. Sci.,* **64,** 241–266.

Spirtes, P., and C. Glymour, 1991: An algorithm for fast recovery of sparse causal graphs. *Soc. Sci. Comput. Rev.,* **9,** 62–72.

——, ——, and R. Scheines, 1991: From probability to causality. *Philos. Stud.,* **64,** 1–36.

——, ——, and ——, 1993: *Causation, Prediction, and Search: Springer Lecture Notes in Statistics*. 1st ed. Springer Verlag, 526 pp.

——, ——, and ——, 2000: *Causation, Prediction, and Search*. 2nd ed. MIT Press, 546 pp.

Steinhaeuser, K., N. V. Chawla, and A. R. Ganguly, 2010: Complex networks in climate science: Progress, opportunities and challenges. *Proc. Conf. on Intelligent Data Understanding,* San Francisco, CA, NASA, 16–26.

Swanson, N., and C. Granger, 1997: Impulse response functions based on a causal approach to residual orthogonalization in vector autoregressions. *J. Amer. Stat. Assoc.,* **92,** 357–367.

Tsonis, A., and P. Roebber, 2004: The architecture of the climate network. *Physica A,* **333,** 497–504.

——, and K. Swanson, 2008: Topology and predictability of El Niño and La Niña networks. *Phys. Rev. Lett.,* **100,** 228502, doi:10.1103/PhysRevLett.100.228502.

——, ——, and P. J. Roebber, 2006: What do networks have to do with climate? *Bull. Amer. Meteor. Soc.,* **87,** 585–596.

——, ——, and S. Kravtsov, 2007: A new dynamical mechanism for major climate shifts. *Geophys. Res. Lett.,* **34,** L13705, doi:10.1029/2007GL030288.

Verma, T., and J. Pearl, 1990: Equivalence and synthesis of causal models. *Proc. Sixth Conf. on Uncertainty in Artificial Intelligence,* Portland, OR, AUAI, 220–227.

Wallace, J., and D. Gutzler, 1981: Teleconnections in the geopotential height field during the Northern Hemisphere winter. *Mon. Wea. Rev.,* **109,** 784–812.

White, H., K. Chalak, and X. Lu, 2011: Linking Granger causality and the Pearl causal model with settable systems. *Proc. Neural Information Processing Systems (NIPS) Mini-Symp. on Causality in Time Series,* Vancouver, British Columbia, Canada, Journal of Machine Learning Research, 1–29.

Woollings, T., B. Hoskins, M. Blackburn, and P. Berrisford, 2008: A new Rossby wave breaking interpretation of the north atlantic oscillation. *J. Atmos. Sci.,* **65,** 609–626.

Wright, S., 1921: Correlation and causation. *J. Agric. Res.,* **20,** 557–585.

——, 1934: The method of path coefficients. *Ann. Math. Stat.,* **5,** 161–215.

Yamasaki, K., A. Gozolchiani, and S. Havlin, 2008: Climate networks around the globe are significantly affected by El Niño. *Phys. Rev. Lett.,* **100,** 228501.

——, ——, and ——, 2009: Climate networks based on phase synchronization track El Niño. *Prog. Theor. Phys.,* **179** (Suppl.), 178–188.