Final Draft
of the original manuscript:

Chrastansky, A.; Callies, U. **:**

Using a Bayesian Network to Summarize Variability in Numerical Long-Term Simulations of a Meteorological–Marine System: Drift Climatology of Assumed Oil Spills in the North Sea

In: Environmental Modeling & Assessment (2010) Springer

DOI: 10.1007/s10666-010-9246-y

Using a Bayesian network to summarize variability in numerical long-term simulations of a meteorological-marine system: drift climatology of assumed oil spills in the North Sea

Alena Chrastansky (corresponding author) & Ulrich Callies
GKSS Research Centre Geesthacht, Germany
Institute for Coastal Research
21502 Geesthacht, Germany
(+49)4152 87 1881
Alena.Chrastansky@gkss.de

Abstract

Climate related scientific analyses of meteorological-marine systems are often based on numerical long-term simulations at high spatial and temporal detail. Such comprehensive data sets require much resources and specific evaluation tools, which sometimes hampers their use within inter-disciplinary projects. In the present study, we propose the use of a Bayesian Network (BN) to represent simulated transports in the North Sea depending on variable external forcing in terms of conditional probabilities. Eliciting probability tables from multi-decadal numerical simulations ensures that all realistic weather and resulting sea state conditions are covered in agreement with the frequency of their occurrence. The probabilistic representation conveniently allows for conditioning numerical simulations on either external forcing (weather conditions) or observed transports. In the latter case, the Bayesian inversion formula becomes involved to transfer information in a direction opposite to causal dependencies encoded in the underlying mechanistic model. We show that simulated travel time distributions even allow for taking into account a substance's specific half-live, although this was not an issue in the original passive tracer simulations.

1 Introduction

Prevailing weather conditions are decisive for the degree to which different coastal stretches are exposed to pollution by oil and other contaminants released on the open sea. For oil spill response planning, for instance, it is important to know characteristic drift paths and times of oil slicks. A thorough contingency analysis, however, needs more than a description of mean conditions. The information basis should include details of variability, the frequency of extreme events, and the coincidence of adverse conditions with regard to wind direction, season, wave heights, and drift times, for instance. A state-of-the-art approach for the provision of such consistent information is the long-term (multi-decadal) reconstruction of past conditions based on numerical model simulations.

Weisse et al. [29] describe the data base coastDat (www.coastdat.de), which compiles model-based reconstructions of atmospheric and oceanic parameters. Realistic numerical simulations of marine currents were produced using wind forcing obtained by dynamic downscaling of re-analyses of the global atmosphere. The detailed information encoded in a vast amount of numerical model outputs far exceeds an aggregate statistical description in terms of means and standard deviations, for instance. Chronic oil pollution along the German North Sea coast is just one application example among many other analyses of recent and possible future changes the data have already been used for [29]. As any oil discharge in the North Sea is prohibited [13]**,** the control of marine oil pollution gains more and more in importance (i.e. [4-5, 12-13, 31]). Despite regular aerial surveillances, however, in many cases illegal washing of oil tanks or discharge of bilge oil goes undetected [26]. Chrastansky et al. [7] used the information from coastDat to support the interpretation of the numbers of beached oil-contaminated sea birds collected along the coasts of Germany as a proxy for a changing general level of pollution and most probable locations of illegal oil spills.

Chrastansky and Callies [6] used the reconstructed hydrodynamic fields from coastDat for an estimate of weather-driven variability of chronic oil pollution along the German North Sea coast. Within a period of several decades, Lagrangian particle drift simulations (particle tracking) were started every 28 hours, each simulation comprising the movements of hundreds of particles released within different source regions. In the model, each particle represents a hypothetical oil spill on the water surface and moves in agreement with North Sea currents and atmospheric wind conditions provided from the coastDat database as function of space and time. The resulting manifold of drift trajectories covers the whole spectrum of possible developments and represents both the most frequent conditions and the occurrence of rare events. Model-based long-term reconstructions of past conditions are nowadays the state-of-the-art approach to consistently generate such information on variability for extended areas (cf. Weisse et al. [29]). The large amount of detailed information, however, comes at a price. Very large data sets necessitate specialized and time-consuming scientific programming for data use and interpretation. This is computationally demanding and becomes particularly tedious when the database must be accessed repeatedly to perform different types of analyses. For many practical applications, however, a full analysis of the raw data is neither feasible nor necessary. Instead, the essence of the detailed long-term data in terms of probability distributions of events and related parameters, including coincidences and causal interactions, would be needed in a more compact format. Users might also wish to link information from the database to information about a specific pollutant's behavior. External information may result from expert knowledge elicitation and be connected with some range of uncertainty.

In this study, we propose a probabilistic data description based on Bayesian Network (BN) technology to conveniently summarizing the essence of the ensemble drift simulations investigated by Chrastansky et al. [6, 7]. In BNs, causal relations are modeled based on probabilistic relationships among the variables of interest [14]. In our example, probability tables are calculated from the vast amount of numerical simulations. The graphical representation of the BN encodes marginal and conditional independence relations that reflect the causal structure underlying these deterministic simulations. Most published applications of BNs in environmental studies rather deal with a combination of expert knowledge, partly formalized in model equations and empirical data. They also often involve presentations of uncertainties that arise from the fact that some relevant pieces of information are missing. Examples are the prediction of fish and wildlife response to land management strategies [19], the investigation of suspended sediment concentrations in alpine catchments as a function of air temperature [21] and the intensive analysis of eutrophication processes using diverse judgment methods [1]. Unlike in these studies, in our example the BN reflects a well-defined causal structure based on physical principles. Instead of filling in missing information, the idea is to properly representing general patterns of interaction within the simulated data. Such aggregated representation may prove itself valuable as an interface between extensive physically based simulations on the one hand and more uncertain consequences on the other. Two previous papers (Chrastansky et al. [6, 7]) related to our example problem illustrate the need for such an interface. We expect that the methodology we present would be adaptable to a range of problems in climate research, for instance.

BNs offer several advantages, although the model construction is challenging and nontrivial [16]. A BN with its intuitive graphical user interface fits the needs of practical decision makers, but may also be useful for training and education purposes. Among others, questions of the following type would be treatable semi-quantitatively without access to the full original database: In which regions are hypothetical pollutions most hazardous for specific coastal stretches? How does a threat depend on the strength of evaporation or other weathering processes implying a shorter half-life of oil or other polluting substances? Are there major seasonal differences? These and other similar questions can be answered by conditioning the BN with regard to certain variables. The BN with a graphical structure encoded in a set of (conditional) probability tables can easily be stored on any PC. It

makes essential information available without a link to the detailed original data sets the BN was derived from. Different software packages, both commercial and free of charge (a list of software may be found i.e. in Korb and Nicholson [17]) allow for the interactive and flexible exploration of a BN's information content and implications.

The paper is structured in the following way: Initially, Section 2 gives an illustration of the hydrodynamic drift simulations (Section 2.1) on which the variables' probability elicitation is based. Section 2.2 describes the atmospheric parameters employed for representing atmospheric forcing in the BN. A general description of BN technology is given in Section 2.3. Construction of our specific BN is explained in Section 3, addressing model structure (Section 3.1) and elicitation of model parameters (Section 3.2) and a concluding summary of the information content of the BN (Section 3.3). Section 4 presents two detailed examples of how the BN can be used for answering specific questions. After a discussion of the BN and its practical utilization (Section 5), conclusions are drawn in Section 6.

2 Material and Methods

2.1 Hydrodynamic drift simulations

The objective of our previous study [6] was to establish a realistic drift climatology of assumed oil spills in the southern North Sea, focusing on chronic oil pollution in the German Bight. As shipping is considered to be the main source of chronic oil pollution [2-3, 8-9, 25], regions where oil spills are supposed to occur were defined along the main shipping lanes. Within each of these source regions, 100 randomly distributed tracer particles representing single hypothetical oil spills were released every 28 hours within the years 1958-2003. Subsequently they were tracked for 60 days by means of a tracer transport model based on simulated winds and 2D currents. Resulting particle trajectories with hourly resolution cover the whole spectrum of possible drift paths including implicit information about their probability of occurrence. For the present study, we confined ourselves to the consideration of 9 source regions (labeled S1-S9 in Figure 1) located within the German Bight. The basic information extracted from each individual numerical simulation consists of a) the numbers of particles from each source region that arrive at 5 different target regions along the German North Sea coast (labeled T1-T5 in Figure 1) and b) the drift-times they need.

FIGURE 1

Figure 1: Particle source and target regions considered in the study. Source regions along the shipping routes are labeled with S1-S9, target regions with T1-T5. Pollutions that originate from more remote sectors of the shipping lanes (light grey shaded) are not taken into account in our study.

Drift simulations were based on pre-calculated hourly hydrodynamic fields stored in the data base coastDat (www.coastdat.de). CoastDat contains high-resolution state-of-the-art re-analyses of past atmospheric and sea state conditions, which have already been employed in various case studies [29]. Two-dimensional hindcasts of marine currents on an unstructured triangular grid, with spatial resolution varying between about 100 m near the coast and a couple of kilometers in offshore regions [24], were produced by running the finite element tide-surge model TELEMAC-2D [11]. At its upper boundary the marine model was forced by hourly NCEP/NCAR re-analyzed wind fields [15], regionalized via dynamical downscaling with the nested regional climate model SN-REMO [20]. Spatial resolution of the atmospheric forcing being used is 50 km.

Wind-induced drift components (1.8 % of the 10 m wind velocity) were superimposed to the movements induced by currents. According to Dick & Soetje [10]**,** this is a proper parameterization

for oil slick movements on the water surface. A random vertical particle motion was included to allow for reduced wind forcing when the tracer particles submerge. Additionally, a random velocity component was used to simulate effects of horizontal diffusion.

2.2 Representation of atmospheric forcing

As simulations are started every 28 hours within the years 1958-2003, Lagrangian tracer particles experience an exhaustive spectrum of realistic time dependent weather conditions. In the BN, however, this complexity of atmospheric forcing cannot be represented. In particular, one must somehow aggregate non-constant weather conditions during each particle cloud's journey. It should be stressed that such simplifications were not used in the numerical drift simulations themselves. For the simplified representation of weather conditions in the BN we used two different approaches.

In our first approach, we dealt with the variables wind direction and wind speed. For each drift simulation we referred to corresponding time series of simulated wind conditions at the island of Heligoland located in the center of the German Bight (54° 11′ N, 7° 53′ E). The strength of relationship between certain wind conditions and simulated coastal pollution was assumed proportional to the time span the wind conditions prevailed. To concentrate on dominant signals, we confined the analysis to the three longest lasting weather conditions within the first three weeks of each individual simulation.

Our second approach borrowed from Chrastansky and Callies [6] is a bit more involved. To identify weather patterns that are most influential for the spatial distribution of coastal pollution, we subjected atmospheric conditions and the resulting outcomes of drift simulations to Canonical Correlation Analysis (CCA) [28]**.** Again, we confined the analysis to atmospheric states during the first three weeks of each particle cloud's drift time, now represented, however, by three consecutive weekly mean Sea Level Pressure (SLP) fields taken from the NCEP/NCAR re-analysis data set [15]. Then each pair of correlated anomaly patterns obtained from CCA (the two most correlated pairs are shown in Figure 2, each pair displayed in one line) consists of a) one pattern of particle advection towards the five target regions and b) one pattern split into three panels related to three consecutive SLP fields.

According to the upper panels in Figure 2, a simultaneous increase (or decrease) of pollution in all coastal regions is strongly correlated ($r = 0.73$) with SLP fields characterized by relatively high (or low) values in the west/southwest and relatively low (or high) values in the east/northeast. Such pressure distributions imply intensified (or weakened) northwesterly wind components associated with increased (or decreased) particle drifts towards the German North Sea coast. Unlike the first anomaly pattern of pollution, the second one (first panel in the bottom row of Figure 2) has opposite signs in the northern and southern regions, respectively. The corresponding SLP anomaly (correlation 0.52) is associated with changing strengths of westerly wind components.

FIGURE 2

Figure 2: Anomaly patterns of SLP and simulated pollution of the German North Sea coast, respectively, as obtained from CCA. In each row, the left panel depicts pollution anomalies that are connected with the triplet of consecutive weekly mean SLP anomalies shown to its right. Correlations are 0.73 and 0.52 for the upper and bottom row, respectively.

Variables included in the BN will be wind direction and wind speed at Heligoland and the CCA time coefficients that represent the changing relevance of the SLP anomaly patterns shown in Figure 2. The CCA time coefficients of SLP are able to explain 32% (first CCA pattern) and 42% (second CCA pattern) of advection variability.

2.3 BN technology

A BN is a probabilistic, directed acyclic graph (DAG) consisting of a set of random variables and a set of directed links between them. Graphically, nodes represent the variables while directed edges encode causal dependencies [16]. In our study, we employ nodes that describe the degrees of freedom each variable has in terms of a number of discrete states which can be either be numbered, labeled or represent intervals. They must, however, always be exhaustive and mutually exclusive [14]. Unless any current evidence exists (e.g. observations) the probability of a variable being in a given state will generally be smaller than one. In case a variable (node) is influenced by other variables, edges from the influencing parent nodes point to the dependent child node [16]. The child node's state probabilities will then be affected by information on the parent nodes' states as specified in a conditional probability table (CPT) associated with the child node.

A most simple example BN made up by only two variables might represent the relationship between the time of the year (season *SN*) and prevailing wind speed (*WS*). The graphical model *SN* → *WS* would correspond with the factorization $P(WS,SN) = P(WS \mid SN)P(SN)$ of the joint probability distribution *P(WS,SN)* [23]. The alternative graphical model *WS* → *SN* corresponding with the factorization $P(WS,SN) = P(SN \mid WS)P(WS)$ would be mathematically equivalent, but specification of the conditional probability $P(WS \mid SN)$ appears to be a more natural option.

Given the CPT $P(WS \mid SN)$, the marginal probability of wind conditions disregarding their seasonal variations can be obtained as

$$P(WS) = \sum_j P(WS \mid SN = sn_j)P(SN = sn_j)$$

with sn_j denoting discrete states of the variable *SN*. Our simple example BN would be made up by the two nodes *WS* and *SN* and the CPT associated with *WS*. In the context of our study, the CPT would represent likelihoods extracted from the results of extensive numerical simulations.

Complete graphs of realistic BNs contain very large numbers of edges, which makes full graphs neither informative nor manageable. Model simplification can be achieved by elimination of edges in the graph. Missing edges, however, imply independence statements that need to agree with either causal reasoning or experience [22]. The kind of implied independence statements depends on the orientations of remaining edges. We will illustrate this point by the inclusion of the atmospheric sea level pressure (*SLP*) as a third variable into the above example.

Figure 3a shows a complete graph with edge directions in agreement with causal reasoning. Three simplified graphs that arise from Figure 3a by omission of one edge at a time are shown in panels (b)-(d) of Figure 3. Implications of the three simplified graphs are profoundly different: Graph (b) states that a priori season (*SN*) is uninformative about air pressure (*SLP*), which is obviously not a valid assumption. Only after evidence about the current state of variable *WS* was entered, any evidence on season (*SN*), for instance, would improve existing knowledge about pressure (*SLP*) (conditional dependence, Kjaerulff and Madsen [16]). Graph (c) states that dependence between the physical variables *WS* and *SLP* can be modeled by changing seasons alone. Given evidence on the current season, any further dependence between *SLP* and *WS* is neglected. This does not correspond with the natural system, in which sea level pressure and wind fields are coupled. Graph (d) retains this physical link between the variables *SLP* and *WS*. At the same time, it allows for a seasonal variation of both of the two variables. The seasonal effect on wind (*WS*), however, is modeled as being channeled through pressure (*SLP*), which means that wind variations being unrelated with pressure show no seasonal dependence. We conclude that graph (d) can be considered as a reasonable model. Likewise, one could choose graph (b) after inversion of the edge connecting *SLP* and *WS*. In either case, information between season (*SN*) and a physical variable (*WS* or *SLP*) at the

end of a serial connection will be transmitted as long as we do not have definite knowledge about the state of the physical variable represented by the middle node.

FIGURE 3

Figure 3: Example of a three node BN: a) complete graph, b)-d) simplified graphs with one edge being discarded.

Using the so-called chain rule (e.g. [22])**,** the factorized trivariat probability distribution reads $P(WS, SLP, SN) = P(WS \mid SLP, SN)P(SLP \mid SN)P(SN)$ according to the complete graph (a) and $P(WS, SLP, SN) = P(WS \mid SLP)P(SLP \mid SN)P(SN)$ according to the simplified graph (d). The effect of graph simplification is that instead of a three-dimensional CPT $P(WS \mid SLP, SN)$ now only the two-dimensional CPT $P(WS \mid SLP)$ needs to be elicited. In our study below, time coefficients of two characteristic atmospheric pressure patterns (cf. section 2.2) will take the part of sea level pressure in Figure 3.

The need for a BN structure that properly mirrors causal relationships has been emphasized by Kjaerulff and Madsen [16]. Another example (Figure 4) extracted from our full study shall illustrate the implications of incorrect independence properties that may be encountered when a BN contains arrows pointing from symptoms to causes. The objective of our study is to represent the risk of coastal oil pollution depending on the locations of hypothetical oil spills and the distribution of prevailing winds. Given that pollution was observed in some target region (*T*), information about past wind conditions (i.e. wind directions (*WD*)) obviously allows for narrowing down the location where the oil spill most probably occurred (source regions (*S*)). Hence, at first sight the structure of graph (a) in Figure 4 seems to be in line with the practical aim to locate possible contaminators.

FIGURE 4

Figure 4: Two BNs that imply different conditional dependences.

The deficiency of graph (a) is, however, that pollution observed in a specific coastal area and knowledge about past wind directions are assumed to be mutually independent pieces of information. Offshore winds in the past would lower our expectation of coastal pollution. Conversely, prevailingness of offshore winds would be disconfirmed by the observation of coastal pollution. Posterior probabilities derived from BN (a) in Figure 4 will therefore be incorrect. For graph (b) with directions of arrows properly representing cause-effect relations, the situation is different. In graph (b), knowledge about the location of an oil spill is assumed non-informative about wind conditions and vice versa. The converging connections in graph (b) allow, however, for a transmission of information between *S* and *WD* whenever evidence on the middle variable target (*T*) is available [16], *S* and *WD* are conditionally dependent. After pollution in some target region was reported, knowledge about past wind directions is informative about possible source regions.

More examples about the functionality and construction of BNs can be found in Jensen [14], Kjaerulff and Madsen [16], Pearl [22-23] and Shipley [27].

3 Model construction

BN model construction always proceeds in two consecutive steps (e.g. Kjaerulff and Madsen [16]): Identification of the probabilistic network's structure and elicitation of model parameters. We used

the software tool Hugin Researcher ™ [18] for model construction and evaluation. Similar tools can be found in Korb and Nicholson [17], for example.

3.1 Specification of BN structure

The first step, structure specification, includes the choice of variables. The two variables that are central for our problem are located on the left of the BN in Figure 5: the partition of total oil pollution among specified source regions (variable 'Source' labeled *S*) and the percentages of simulated passive tracer particles that arrive in specified target regions (variable 'Target' labeled *T*). Note that the overall amount of pollutant discharge from all source regions is assumed to stay on an unspecified constant level. Both of the nodes *S* and *T* are chance nodes that represent random or uncertain variables.

The drift model underlying our BN assumes that the water pollutant behaves like a passive tracer. This is obviously not realistic. The chance node 'Drift Time' (*DT*), however, implemented as a child of *S* and *T*, provides a probability distribution of particle travel times. This information allows the user to re-weight passive tracer advection according to some specified half-life time τ. Assuming exponential decay processes, modified estimates (*T**) of the pollutant's arrival rate in various target regions can be calculated. Note that specification of τ was not an issue in the underlying tracer simulations. Therefore, the node τ establishes an interface between the pre-calculated transport climatology and specific information in the context of practical applications.

Motivated by the study of Chrastansky and Callies [6], we introduced time coefficients of CCA patterns for SLP, obtained from correlating SLP fields with simulated particle advection rates (cf. Section 2.2), as the primary representatives of atmospheric forcing. The nodes *SLP 1* and *SLP 2* correspond with the time coefficients of the two SLP anomaly patterns shown in Figure 2.

Additional nodes 'Wind Speed' (*WS*) and 'Wind Direction' (*WD*) correspond with local conditions at Heligoland (see section 2.2). According to the BN in Figure 5, however, acquiring evidence on local winds does not modify estimates of particle advection or corresponding travel times (conditional independence of *T* and *WS*, for instance) given that evidence for both *SLP 1* and *SLP 2* already exists. Another consequence of the selected BN structure is that *WS* and *WD* are conditionally independent given values of *SLP 1* and *SLP 2*. This assumption appears justified if the CCA time coefficients are proper surrogates for large-scale wind fields and local wind speed components do not much depend on wind direction.

Entering evidence for nodes in the BN's right hand side subdomain (*SN*, *SLP1*, *SLP2*, *WS*, *WD*) enables a user to study both particle advection and corresponding drift times as functions of prevailing wind conditions, for instance. Alternatively, a user might be interested in seasonal variations of coastal pollution. In keeping with causal relationships, information on the choice of a specific season (*SN*) is always transmitted through changing probability distributions of CCA time coefficients for SLP.

We conclude this section with a short summary of the BN's structure (Figure 5). At its root there is node *S*, in which different source regions for oil spills may be selected. Considering a maximum integration period of 60 days, any oil spill will either hit some sector of the German North Sea shoreline (states in node *T*) or move elsewhere (state 'none' of *T*). For each pair of source and target regions a characteristic distribution of drift times (*DT*) is obtained from the underlying ensemble of numerical simulations. Distributions of both variables, *T* and *DT*, however, will change when focusing on certain weather conditions represented by nodes in the right hand side subdomain (see also discussion of Figure 4 in Section 2.3). Sea level pressure (*SLP1* and *SLP2*) is modeled as a function of season (*SN*). The influence of *SN* on wind conditions (*WD* and *WS*) is assumed to be channeled through SLP patterns. The nodes *SLP1* and *SLP2* remain unconnected as the CCA time

coefficients they represent are uncorrelated by construction. Finally, *T** is a duplicate of *T*, introduced to allow for re-weighting the distribution of *T* according to drift time (*DT*) in case that a finite half-life (τ) was specified. *T* always refers to passive tracers, regardless of the choice of τ.

3.2 Definition of states and elicitation of model parameters

For each node in the BN a set of discrete states must be defined the corresponding variable may attain. For source node *S* these states are labeled S1-S9 according to the nine source regions tracer particles may be released from (cf. Figure 1). States of target nodes *T* and *T** refer to the five target regions (T1 - T5) shown in Figure 1. Additional states 'none' are applicable when oil spills do not affect the German coast.

Travel time *DT* is resolved with 6 intervals of non-uniform lengths (0-5 days, 5-10 days, 10-15 days, 15-20 days, 20-30 days and 30-60 days). An additional state 'infinite' covers the case that no oil hits the German coast within the maximum simulation time of 60 days. This state corresponds with the state 'none' in variable *T*. Possible values of half-life τ are 5, 10, 20, 30 and 50 days.

States of the nodes *SLP 1* and *SLP 2* were defined in terms of multiples of the standard deviation of the corresponding time series (note that CCA time coefficients are related to SLP anomalies and therefore have zero means): State 'o' comprises values within +/- 0.5, state '+' ('-') values between 0.5 and 1.5 (-0.5 and -1.5), and state '++' ('--') values that exceed 1.5 (-1.5) standard deviations. Wind direction *WD* is resolved by the eight states SW, W, NW, N, NE, E, SE and S. States 2-8 of wind speed *WS* follow the Beaufort (bft) classification [30]. The season node *SN* differentiates between four seasons 'spring' (Mar-May), 'summer' (Jun-Aug), 'fall' (Sep-Nov), and 'winter' (Dec-Feb).

With regard to possible source regions (chance node *S*) we assumed a non-informative uniform prior distribution. Alternatively, we might have chosen a prior distribution obtained from German aerial surveillance data [5, 31].

For chance nodes *T* and *DT*, both marginal probabilities and conditional probability tables (CPTs) were elicited from pre-calculated drift simulations and corresponding atmospheric forcing. For this purpose, extensive data tables made up by values for the node itself and each of its parent nodes were imported into the software tools of Hugin Researcher ™ [18].

Marginal probabilities of the node SN reflect the number of simulations that were initialized in different seasons of the year. The distribution is very close to uniform.

For specification of the CPTs for *WS* and *WD* the procedure based on data tables was slightly extended. Following the approaches described in Section 2.2, each numerical drift simulation is associated with sharp values for *SLP 1* and *SLP 2* but triples of values for *WS* and *WD*. Weighting factors reflect relative frequencies of occurrence of three different wind conditions during a given drift simulation. To deal with this type of information, we replicated samples in the data table with replacement of the wind related values. Different numbers of auxiliary samples with identical wind values were introduced in accordance with the relative importance of each of the three wind conditions.

Parameter elicitation for the re-scaled target node *T** needs no recourse to the database of numerical simulations. Instead, a simple data table was generated based on an exponential decay formula.

For the marginal node τ, the basic assumption is a non-informative (i.e. uniform) prior distribution. Given information on both *T* and *T**, for instance, this prior would be changed into a posterior

estimate of τ (conditional dependence of *T* and τ given *T**). In the examples discussed below, however, we will always assume a fixed value for τ (cf. Figure 5, for instance).

3.3 The BN's information content and validity

The BN summarizes dominant patterns of dependence between variables in an already existing database of extensive numerical drift simulations with high resolution in both space and time (cf. Section 2.1). It is important to note that probability distributions in the BN do not reflect inaccuracies of the underlying numerical simulations. Instead, the majority of chance nodes represents the spectrum of weather conditions and corresponding implications for drift paths and times within a 46-year period. Exceptions are the two nodes *S* (source regions) and τ (the substance's half-life) which represent external assumptions to be specified in user defined scenarios.

Validation of the hydrodynamic drift simulations already discussed in two previous studies ([6] and [7]) is not subject of consideration in the present paper. Chrastansky et al. [7]) showed that variations in the annual mean numbers of oil-contaminated bird corpses beached along the German North Sea shoreline correspond reasonably well with modeled advection rates. We take that as an indicator that the model provides a useful description of drift behavior in the German Bight. Generally, validation of long distance drift simulations is difficult as relevant data are rare.

The structure of a BN mirroring the outcomes of deterministic simulations does not need sophisticated techniques of expert judgment. In this regard, the situation in this paper differs substantially from dealing with systems where parameter interactions are unclear. Hydrodynamic currents and resulting drift paths depend on time dependent atmospheric wind fields, and conditional independence relationships encoded in the graph must be consistent with such cause effect relationships (cf. Section 2.3). Generally, the description would become much more uncertain, if non-conservative tracers with a complex behavior were included. We confined ourselves, however, to a treatment of simple tracer particles with a user-defined half-life. Again coupling of this half-life to the BN follows logical rules, considering that simulated travel times are available from the numerical simulations.

The graph in Figure 5 involves, however, also substantial simplifications that have effects on calculated conditional (not marginal!) probabilities. Just two SLP patterns are obviously insufficient to fully resolving atmospheric variability including seasonal cycles. Hence, a BN completed by additional links SN→T and SN→DT that allow for information transfer from season SN to T and DT not being channeled through the two SLP nodes would improve the representation of seasonal effects. Table 1 compares conditional probabilities obtained from such a complete BN with those obtained from the simplified version in Figure 5, assuming that variable SN is in the state 'summer'. In spite of clear differences, the general patterns of change with regard to the unconditional distributions in Figure 5 are similar. This indicates that the two SLP patterns indeed cover main effects of seasonal variability. The disadvantage of using the BN with additional links is that it improves the representation in a merely descriptive way that does not allow for explaining effects in terms of changing atmospheric forcing.

In the following examples we will adhere to the simplified graph shown in Figure 5.

Table 1: Conditional probability distributions of *T* and *DT* for *SN*=summer. Comparing values from a complete BN containing all edges and the simplified BN shown in Figure 5.

T			DT		
states	complete BN	simplified BN	states	complete BN	simplified BN
none	27	34	infinite	27	34
T1	17	16	]00,05[	2	2
T2	23	19	[05,10[	8	8
T3	14	12	[10,15[	11	10
T4	13	12	[15,20[	11	10
T5	6	6	[20,30[	18	16
			[30,60]	23	20

4 Two example studies using the BN

FIGURE 5

Figure 5: A BN to represent the climatology and spatial distribution of weather driven coastal pollution from assumed oil spills. Probability distributions shown result from the assumption of a pollutant's half-life of τ=20 days.

Figure 5 shows the basic state of the BN. Bar charts represent probability distributions for all variables in terms of percentages. Most distributions are marginal distributions of variables in the underlying numerical simulation results. The only exceptions are an assumed half-life of τ = 20 days and the resulting posterior distribution of *T** conditioned by this value. Unless any evidence is entered for other variables, *T** remains the only variable affected by the specification of τ.

According to the distribution of *T*, about 46% of passive tracer particles from any source region do not reach the coast within the simulation period of 60 days. For the remaining 54% of passive tracer particles drift times vary substantially. In most cases the assumed half-life τ = 20 days is exceeded. Hence, when the passive tracer assumption is relaxed, even about 80% (instead of 46%) of the pollutant are estimated to not hitting the shoreline (cf. node *T**). All percentages mentioned are averages over the whole spectrum of possible weather conditions throughout the year (season *SN* remains unspecified).

4.1 Example 1: Effects of releases from source region S4

Assume now that we are interested in the consequences of an oil spill in source region S4. S4 in the interior German Bight is located close to the shoreline, so that the percentage of passive tracer particles that hit the German coast increases from 54% for unspecified source regions (cf. *T* in Figure 5) to about 76%. In particular for target regions T2-T4 pollution increases, whereas regions T1 and T5 are found to be less threatened than on average by all source regions (not shown).

Probabilities for coastal pollution may substantially vary for different seasons. Referring again to hypothetical releases in source region S4, panels a and b in Figure 6 show posterior distributions of *T* for summer and winter, respectively. The overall threat of coastal pollution along the German North Sea coast is more pronounced in summer (88%) than in winter (68%). For target regions T3 and T4, the importance of summer is most pronounced. By contrast, target region T1 experiences an opposite seasonal trend with a maximum risk in winter (21%) and a minimum risk in summer (9%). The differences can be attributed to different orientations of coastal stretches relative to prevailing wind directions. In fall and in winter westerly and southwesterly wind component preponderate. In spring and summer, on the other hand, southerly winds become less and northerly winds more frequent.

FIGURE 6

Figure 6: Conditional probability distributions for target region (*T*) and wind direction (*WD*) assuming that a hypothetical oil spill took place in source region S4. Additional assumptions on season or atmospheric pressure patterns are listed in the headers of individual panels.

This dependence is well illustrated by conditioning target variable *T* on pressure pattern *SLP 1*. (while setting SLP 2 to its neutral value 'o'). Observing *SLP 1* = '+' makes summer the most probable season (55%, not shown). According to Figure 6c, however, the focus of coastal pollution on target regions T2, T3 and T4 becomes even more pronounced than for the mean summer conditions (Figure 6a), due to enhanced north-westerly winds (cf. Figure 6d). On the other hand such wind conditions shield target region T1 from pollution from the more southern source region S4. The overall probability that under the assumed weather conditions oil from region S4 reaches any part of the German North Sea coast is 98% assuming passive tracers, or 50% and 13% assuming a half-life of 20 and 5 days, respectively (not shown).

Choosing instead state '-' for *SLP 1* makes summer season very unlikely (only about 6%), the most probable seasons are now winter and fall (not shown). Southerly winds (cf. Figure 6f) imply the highest risks for target T1 (cf. Figure 6e), while risks for the southern target regions T3-T5 are very low. The situation resembles the winter situation shown in Figure 6b. The overall threat for the German coast is much smaller than for *SLP 1* = '+' (44% assuming passive tracers; 15% assuming a substance half-life of 20 days) as travel times are clearly longer (not shown).

4.2 Example 2: Risk of pollution in coastal area T3

The numerical simulations underlying our BN are made up of trajectories integrated forward in time. For addressing the question of to which extent different source regions of a hypothetical pollutant pose a threat to particular coastal stretches, backward trajectory simulations would be an option. In a probabilistic framework, however, the Bayesian inversion formula allows for deriving the same type of information also from ensembles of forward simulations.

FIGURE 7

Figure 7: The BN with evidence for coastal pollution in target region T3 and a pollutant's half-life τ=20d.

The probability of particle advection towards some specific target region depends on both the source region, where the oil spill takes place, and prevailing wind conditions. Figure 7 shows how the selection of *T** = T3 produces a non-uniform (conditional) probability distribution for possible sources of the pollution. Instantiation of *T** instead of *T* brings the assumed substance half-life of 20 days into play. Due to conditional dependence of *DT* and τ given *T**, the distribution of drift times shifts towards values much smaller than when evidence was provided for passive tracer particles (i.e. for *T* instead of *T**, not shown). The probability of coastal pollution arising from the nearby source region S4 increases from 22% for passive tracers to 30% (Figure 7) for τ = 20d. Choosing a substance half-life of only 5 days raises the probability of pollution stemming from source S4 to even 47% (not shown). Instantiation of *T** (i.e. assuming that coastal pollution was observed) also affects probability distributions of meteorological variables. According to Figure 7, the posterior probability of *SLP 1* being in state '+', for instance, is 54% (unconditional value is 28%).

Another important aspect is that specification of target regions makes source regions and atmospheric conditions conditionally dependent. Given evidence on pollution in any target region,

wind conditions become informative about the distribution of possible source regions and vice versa. Assume that we are interested in exploring possible impacts of the distant source region S1 on target region T3. Figure 8 shows a selection of panels from Figure 7 that substantially change when setting *S* to state S1. The two lines of panels in Figure 8 differ with regard to the choice of τ.

FIGURE 8

Figure 8: Conditional probabilities for drift time (*DT*), season (*SN*) and pressure field (*SLP 2*). First line: Given pollution in coastal region T3, substance half-live τ = 20 days, oil released in source S1. Second line: As before but for τ = 20.

Choosing *S*=S1 (and keeping τ = 20d) shifts the distribution of drift times to higher values (cf. Figure 7 and Figure 8a), which is due to the relatively long distance to be covered between source S1 and target T3. The corresponding drift processes primarily occur for states 'o' or '+' of node *SLP 2*. A smaller substance half-life of only 5 days (Figure 8b) makes possible drift times shrink, probabilities of states '+' and even '++' of node *SLP 2* further increase. The preferred season for such atmospheric conditions associated with strong westerly winds is now clearly winter (probability 42%) instead of summer in the event that source regions are unspecified (Figure 7).

5 Discussion

The BN we discussed represents the essentials of numerical Lagrangian drift simulations. The underlying mechanistic model made it easy to define the BN's structure in line with causal reasoning. Parameter elicitation, the second step of model construction, could be based on a large ensemble of numerical simulations covering the full spectrum of realistic weather conditions that occurred during a time span of several decades.

The probabilistic representation in a BN allows for conditioning variables on either weather conditions or the resulting simulated coastal pollution, for instance. In the latter case, the Bayesian inversion formula is the basis for information transfer in directions opposite to causal dependences. Different software packages (e.g. Korb and Nicholson [17]) allow for fast propagation of any combination of evidence throughout the whole network.

Our numerical simulations were confined to advective transports. Corresponding nodes in the basic BN may be linked, however, to nodes that represent parameters not treated in the simulations. We chose substance half-life τ as an example. Another example we explored (not shown) was a wind dependent probability for illegal oil dumping, taking into account that contaminators supposedly try to remain undetected. It turned out, however, that conditioning on wind speed at the time of discharge did not substantially change the climatology of subsequent drifts and resulting coastal pollution. The two examples illustrate how sensitivities with regard to uncertain parameters may be estimated from a BN without repeating time-consuming numerical simulations.

There are many different ways our prototypical network could be modified or extended for other applications. Weathering processes now represented by the half-life τ might be modeled in more detail as functions of prevailing weather conditions. For the hydrodynamic simulations, we assumed movements induced by currents to be superimposed by an extra drift component amounting to 1.8% of the 10m wind velocity. In a more general BN, wind drift might be established as an additional parent node of *T* and *DT*. Re-calibration of conditional probability tables for the latter two nodes would need, however, extended numerical ensemble simulations.

The season node *SN* in the network provides an important interface for the inclusion of biologically oriented aspects. Vulnerability of a bird species depends on its habitat and on its seasonal molding and breeding behavior. One may ask to which degree, under which circumstances, and from which primary sources of pollution a given bird species is most endangered. Chrastansky et al. [7] showed that the interpretation of beached bird survey programs benefits from detailed numerical drift simulations.

Another type of observation to be combined with the results of numerical modeling is aerial surveillance data. In Section 4.2, prior probabilities for oil discharge were assumed to be the same in each source region (cf. Figure 5). Accordingly, the posterior distribution for *S* (Figure 7) given pollution in target region T3 should be read as the sensitivity of region T3 with regard to hypothetical pollutions in different source regions. A more realistic estimate of the posterior probability that a given pollution stems from a certain source region could be obtained when data from aerial surveillances were introduced as prior probability distributions.

6 Conclusions

Long-term simulations with process-based numerical models have become a state-of-the-art tool for the assessment of changing natural environments. In climate research, complex models are used for both the reconstruction of past climate variability and the construction of possible future scenarios. Model-based reconstructions providing kind of laboratory for risk-assessment studies have also been used as a surrogate for natural conditions [29].

Traditionally, the essentials of long-term simulations and corresponding scientific analyses are made available in the form of written papers. Such static presentations naturally focus on a specific perspective chosen by the author. A modified view at the simulated data will usually need direct access to the full database. For many studies, e.g. studies with a wider interdisciplinary scope, this will often be beyond the means. Nevertheless, even for the design of more in-depth analyses it is often desirable to get a quick overview of relevant features and relations represented in comprehensive simulations. Our study has shown how the technique of Bayesian networks might be employed to meet such demands.

In a first step, a BN will always reflect its designer's scientific interests, focusing on specific variables or aspects of co-variability in a data set. Our example illustrated, however, the modular structure a BN representing numerical simulations usually will have. Elicitation of the conditional probability table for any given node remains local in the sense that data tables just need to contain information on those variables that directly influence the node of interest. As a result, an existing BN can often be extended without recalibration of most of its already existing components.

We introduced a half-life τ of the drifting substance to give a practical example of how generic passive tracer simulations under realistic weather conditions can be linked with a specific substance's properties. This very efficient approach avoids repetition of computationally demanding numerical simulations and might be useful particularly in the context of contingency planning where substance properties are only vaguely defined. Another issue would be the combination of numerical simulations with monitoring data, e.g. by using aerial surveillance data for defining the prior probabilities for sources of pollution (node *S*). A BN, however, might also encompass more qualitative expert knowledge.

A crucial problem to be solved for a successful representation of detailed simulations in a BN is the reduction of dimensionality. In the present study, we utilized the results of canonical correlation analysis adapted from a previous study for a simplified description of weather conditions in terms

of weekly mean sea level pressure patterns. One might also use different techniques, employ different variables, and refer to different scales in space or time. The BN approach we proposed for the description of output from large numerical simulation models, however, does not substitute the use of common multivariate statistical tools.

We expect a BN fitted to an existing set of long-term numerical simulations to be subject to permanent development. Extensions and improvements of the BN may arise from new scientific analyses and subsequent applications of the data. On the other hand, an evermore-detailed BN might inspire the conception of new investigations or help to promote the database. The BN's intuitive graphical presentation helps to make a database accessible even for non-scientific users. Furthermore, the data representation based on the BN technology is fast and flexible enough to be run in the background of web applications. BN manipulations based on libraries written in Fortran, C or Java can be blended with common internet programming languages. For a more intuitive display of calculated probabilities and regions they refer to, the BN output might be combined with GIS items such as geographical maps.

References

1. Borsuk, M.E., Stow, C.A, Reckhow, K.H. (2004). A Bayesian network of eutrophication models for synthesis, prediction, and uncertainty analysis. *Ecological Modelling*, 173, 219-239.
2. Camphuysen, C.J. (1998). Beached Bird Surveys Indicate Decline in Chronic Oil Pollution in the North Sea. *Marine Pollution Bulletin*, 36 (7), 519-526.
3. Camphuysen, C.J. (2007). Chronic Oil Pollution in Europe. A status report. Royal Netherlands Institute for Sea Research, commissioned by International Fund for Animal Welfare, Brussels.
4. Camphuysen, C.J., Heubeck, M. (2001). Marine oil pollution and beached bird surveys: the development of a sensitive monitoring instrument. *Environmental Pollution*, 112, 433-461.
5. Carpenter, A. (2007). The Bonn Agreement Aerial Surveillance programme: trends in North Sea oil pollution 1986-2004. *Marine Pollution Bulletin*, 54,149-163.
6. Chrastansky, A., Callies, U. (2009). Model-based long-term reconstruction of weather-driven variations in chronic oil pollution along the German North Sea coast. *Marine Pollution Bulletin*, doi:10.1016/j.marpolbul.2009.03.009.
7. Chrastansky, A., Callies, U., Fleet, D.M. (2009). Estimation of the impact of prevailing weather conditions on the occurrence of oil-contaminated dead birds on the German North Sea coast. *Environmental Pollution*, 157, 194-198.
8. Dahlmann, G. (1985). Herkunft der Ölverschmutzung der Nordsee durch Öl und Schiffsmüll. Berlin: Umweltbundesamt, 34-55.
9. Dahlmann, G., Timm, D., Averbeck, C., Camphuysen, C., Skov, H., Durinck, J. (1994). Oiled Seabirds - Comparative Investigations on Oiled Seabirds and Oiled Beaches in the Netherlands, Denmark and Germany (1990-1993). *Marine Pollution Bulletin* 28(5), 305-310.
10. Dick, S., Soetje, K.C. (1988). Ein numerisches Modellsystem zur Vorhersage der Drift und Ausbreitung von Öl in der Deutschen Bucht. Umweltforschungsplan des Bundesministeriums für Umwelt, Naturschutz und Reaktorsicherheit, Hamburg: Forschungsbericht 102 03 216.
11. Hervouet, J.M. van Haren, L. (1996). TELEMAC2D version 3.0 Principle Note. Chatou CEDEX. Rapport EDF HE-4394052B.
12. International Maritime Organization (1982). Regulations for the Prevention of Pollution by Oil. London: International Maritime Organization.

13. International Maritime Organization (2002). International Convention for the Prevention of Pollution from Ships, 1973, as modified by the Protocol of 1978 relating thereto (MARPOL). http://www.imo.org/TCD/contents.asp?doc_id=678&topic_id=258#19. Accessed 6 May 2009.
14. Jensen, F.V. (1996). *An introduction to Bayesian Networks*. London: UCL Press.
15. Kistler, R. et al. (2001). The NCEP-NCAR 50-year reanalysis: monthly means CD-ROM and documentation. *Bulletin of the American Meteorological Society*, 82(2), 247-268.
16. Kjaerulff, U.B., Madsen, A.L. (2008). *Bayesian Networks and Influence Diagrams*. New York: Springer.
17. Korb, K.B., Nicholson, A.E (2003). *Bayesian artificial intelligence*. Boca Raton: CRC Press.
18. Madsen, A.L., Jensen, F., Kjaerulff, U.B., Lang, M. (2005). The HUGIN tool for probabilistic graphical models. *International Journal on Artificial Intelligence Tools*, 14(3), 507-543.
19. Marcot, B.G., Holthausen, R.S., Raphael, M.G., Rowland, M.M., Wisdom, M.J. (2001). Using Bayesian belief networks to evaluate fish and wildlife population viability under land management alternatives from an environmental impact statement. *Forest Ecology and Management*, 153, 29-42.
20. Meinke, I., von Storch, H., Feser, F. (2004). A validation of the cloud parameterization in the regional model SN-REMO. *Journal of Geophysical Research*, 109, D13205, doi:10.1029/2004JD004520.
21. Mount, N., Stott, T. (2008). A discrete Bayesian network to investigate suspended sediment concentrations in an Alpine proglacial zone. *Hydrological Processes*, 22, 3772-3784.
22. Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems – Networks of Plausible Inference*. San Francisco: Morgan Kaufmann Publishers.
23. Pearl, J. (2000). *Causality. Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.
24. Plüß, A. (2004). Das Nordseemodell der BAW zur Simulation der Tide in der Deutschen Bucht. *Die Küste*, 67, 83-127.
25. Reineking, B. (2005). Harbors and shipping. Wadden Sea Quality Status Report 2004. Common Wadden Sea Secretariat. Wilhelmshaven: Wadden Sea Ecosystem, 19, 35-38.
26. Schallier, R., Lahousse, L., Jacques, T.G. (1996). Surveillance aérienne : pollutions marines causées par les navires dans la Zone d' Intérêt de la Belgique en Mer du Nord. Rapport d' activité 1991-1995. Unité de Gestion du Modèle Mathématique de la mer de Nord. Bruxelles.
27. Shipley, B. (2001). *Cause and Correlation in Biology. A User's Guide to Path Analysis, Structural Equations and Causal Inference*. Cambridge: Cambridge University Press.
28. Storch von, H., Zwiers, F.W. (1999). *Statistical Analysis in Climate Research*. Cambridge: Cambridge University Press.
29. Weisse, R., Storch von, H., Callies, U., Chrastansky, A., Feser, F., Grabemann, I., Guenther, H., Pluess, A., Stoye, T., Tellkamp, J., Winterfeldt, J., Woth, K. (2009). Regional meteo-marine reanalyses and climate change projections: Results for Northern Europe and potentials for coastal and offshore applications. *Bulletin of the American Meteorological Society*, doi:10.1175/2008/BAMS2713.1.
30. World Meteorological Organization (2008). Guide to Meteorological Instruments and Methods of Observation. WMO-No. 8, Seventh edition (page 439).
31. Viebahn von, C. (2001). Oil Spill Statistics and Oil Spill Monitoring. Hamburg: DGMK-Research Report, 564.

Figure 1

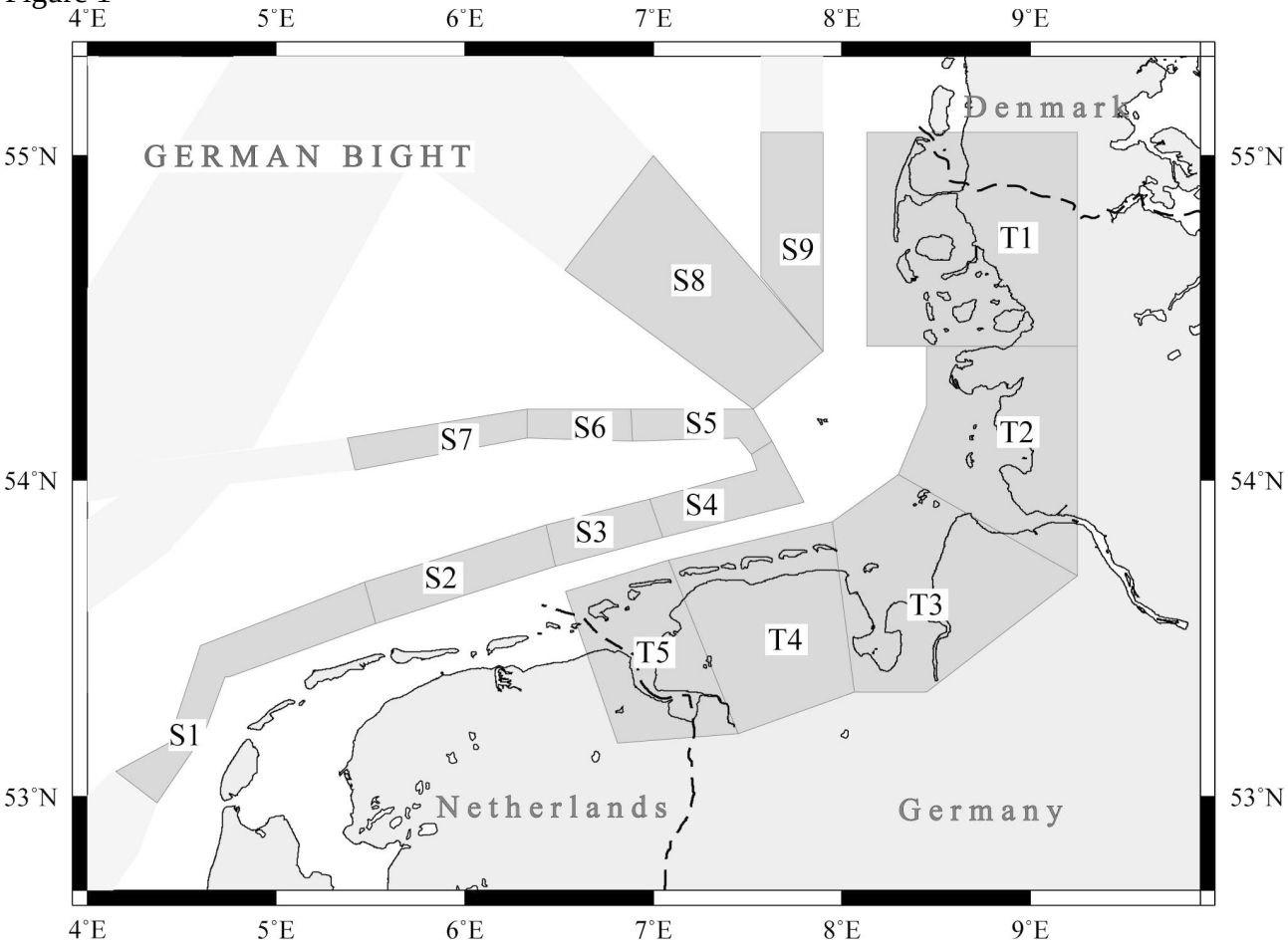


Figure 2

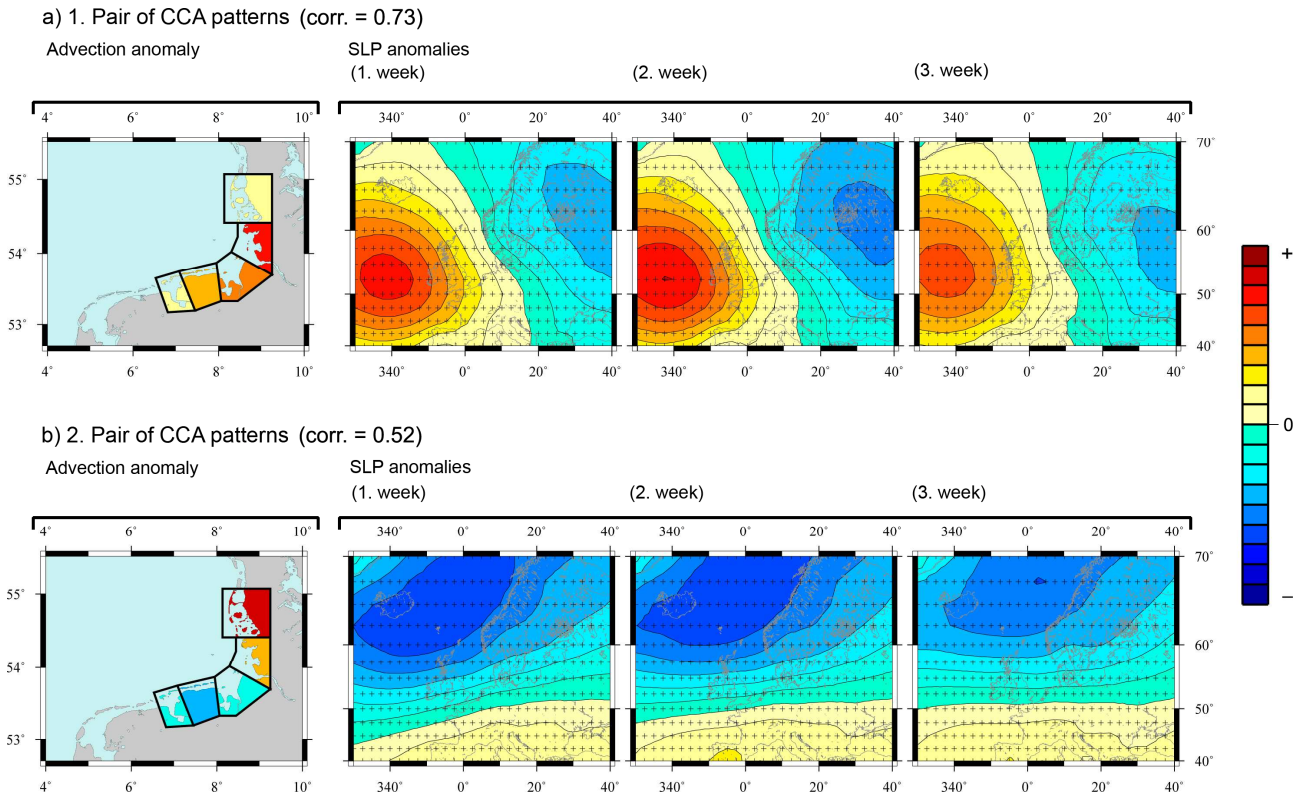


Figure 3

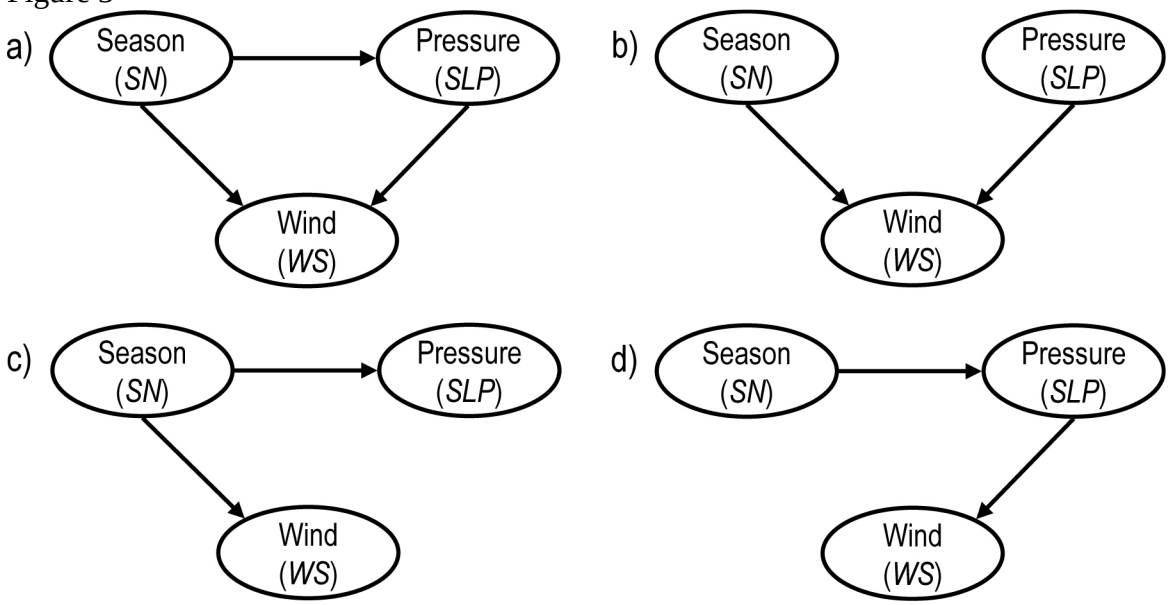


Figure 4

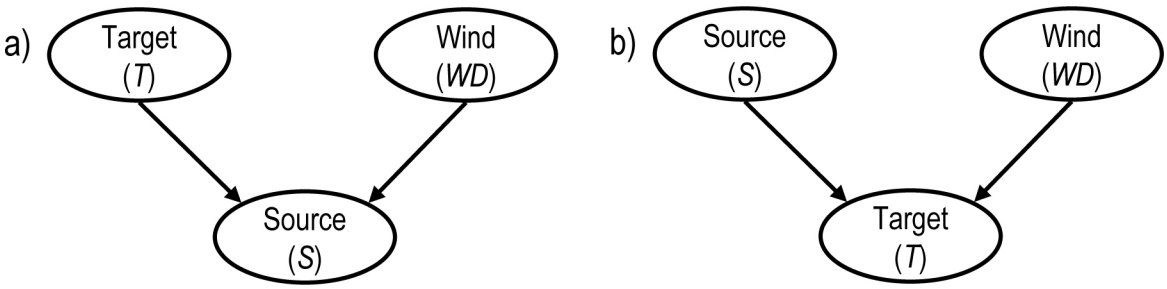


Figure 5

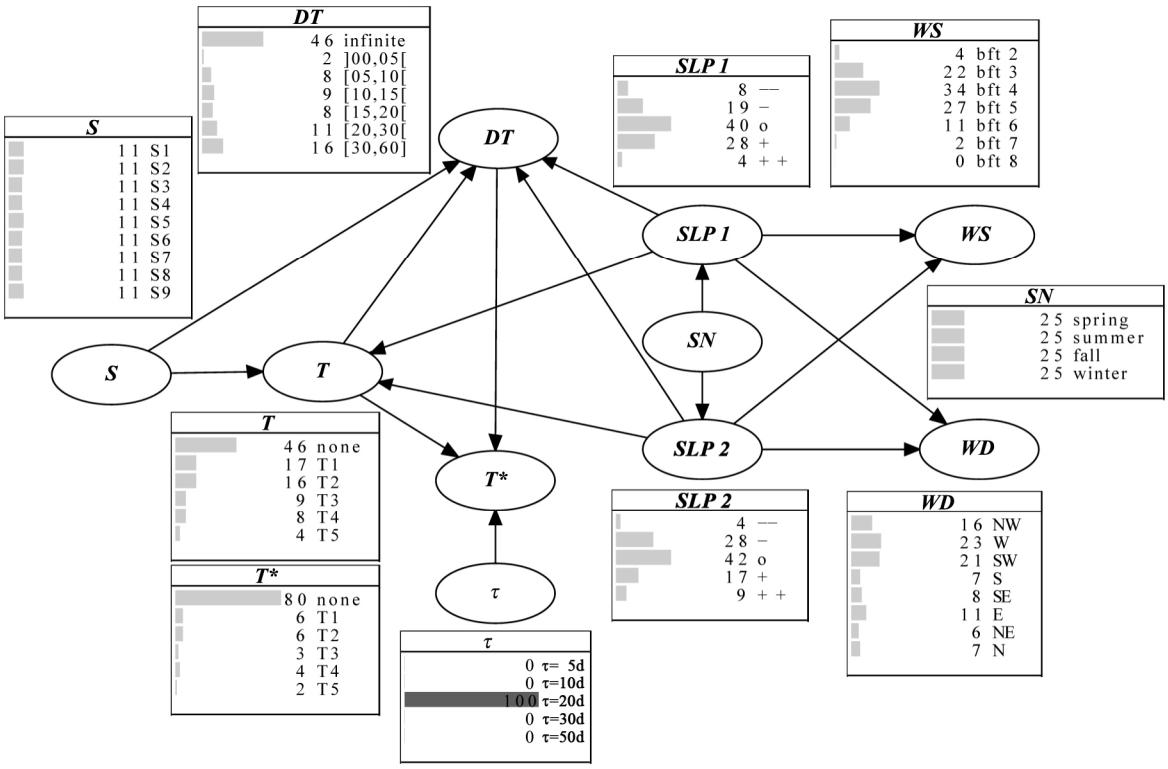
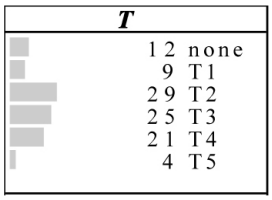
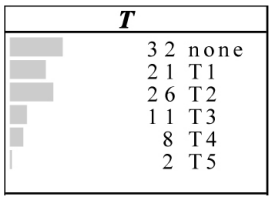


Figure 6

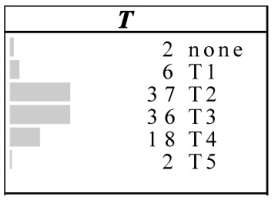
a) *S*=S4, *SN*=summer



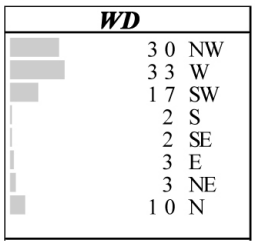
b) *S*=S4, *SN*=winter



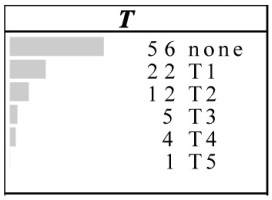
c) *S*=S4, *SLP 1*=„+“, *SLP 2*=„o“



d) *S*=S4, *SLP 1*=„+“, *SLP 2*=„o“



e) *S*=S4, *SLP 1*=„-“, *SLP 2*=„o“



f) *S*=S4, *SLP 1*=„-“, *SLP 2*=„o“

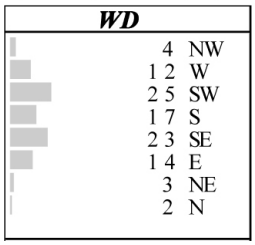


Figure 7

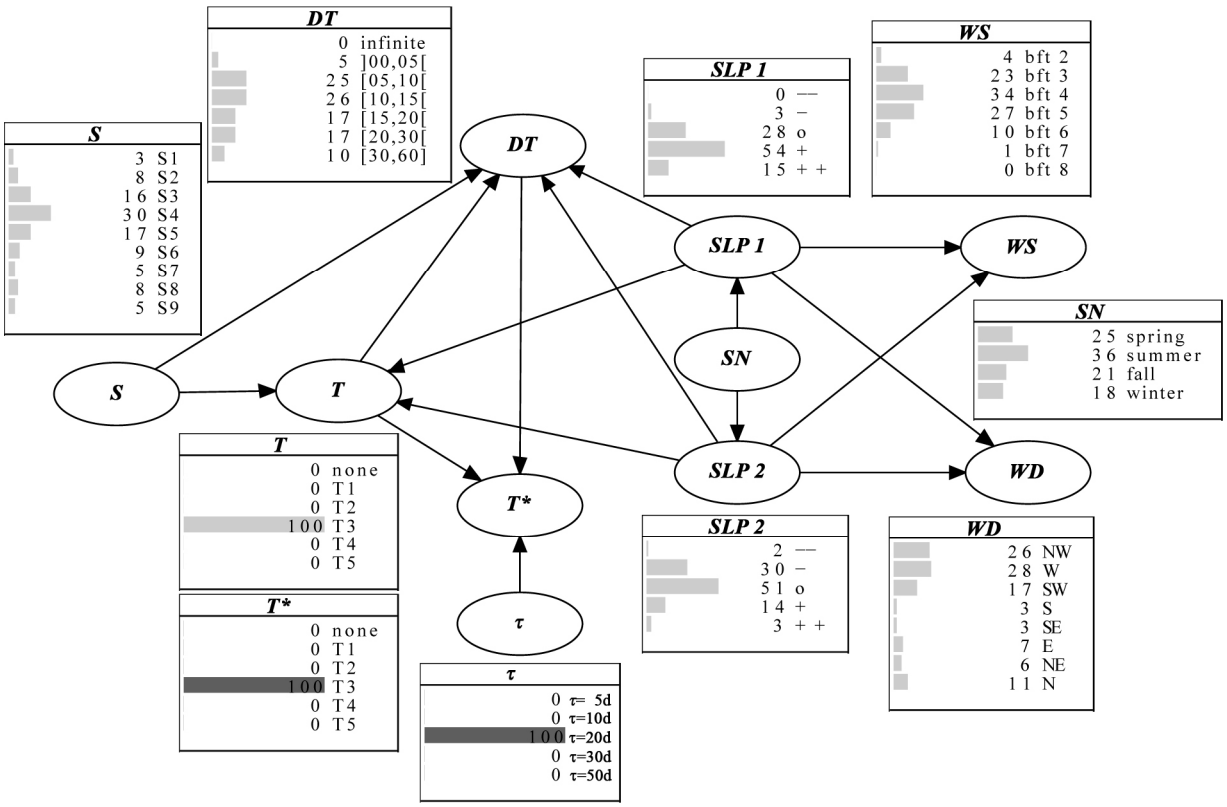
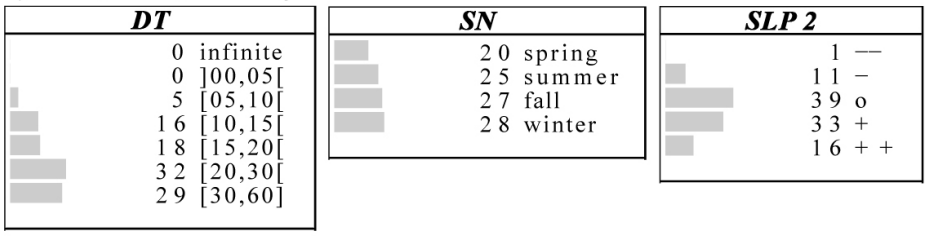


Figure 8

a) *S*=S1, *T*=T3, *τ*=20 days



b) *S*=S1, *T*=T3, *τ*=5 days