NIASRA
NATIONAL INSTITUTE FOR APPLIED
STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

University of Wollongong, Australia

Working Paper

06-20

Measuring, Mapping, and Uncertainty Quantification in the Space-Time Cube

Noel Cressie and Christopher K. Wikle

National Institute for Applied Statistics Research Australia, University of Wollongong,
Wollongong NSW 2522, Australia Phone +61 2 4221 5076, Fax +61 2 4221 4998.
Email: karink@uow.edu.au

Measuring, mapping, and uncertainty quantification in the space-time cube

Noel Cressie, University of Wollongong, Australia
Christopher K. Wikle, University of Missouri, USA

Abstract

The space-time cube is not a cube of course, but the idea of one is useful. Its base is a spatial domain, D_t, and the "cube" is traced out by a process of spatial domains, $\{D_t : t \geq 0\}$. Now fill the cube with a spatio-temporal stochastic process $\{Y_t(\mathbf{s}) : \mathbf{s} \in D_t, t \geq 0\}$. Assume that $\{D_t\}$ is fixed and known (but clearly it too could be stochastic). Slicing the cube laterally for a fixed t_0 generates a spatial stochastic process $\{Y_{t_0}(\mathbf{s}) : \mathbf{s} \in D_{t_0}\}$. Slicing the cube longitudinally for a fixed $\mathbf{s}_0$ generates a temporal process $\{Y_t(\mathbf{s}_0) : t \geq 0\}$ that, after dicing, yields a time series, $\{Y_0(\mathbf{s}_0), Y_1(\mathbf{s}_0), \ldots\}$. These are the main highways that traverse the cube but other, less-traveled paths, can be taken. In this paper, we discuss spatio-temporal data and processes whose domain is the space-time cube, and we incorporate them into hierarchical statistical models for spatio-temporal data.

Introduction

This paper gives an philosophical and etiological discussion of spatio-temporal statistics. We draw on ideas presented in our two recent books on the topic (Cressie and Wikle, 2011; Wikle et al., 2019).

Causation is the "holy grail" of Science, and hence to infer cause-effect relationships (i.e., "why") it is essential to keep track of "when," since a cause always precedes an effect. Knowing "where" recognizes the importance of knowing the "lie of the land" and the multi-dimensional world in which we live. Hence, in order to answer the "why" question, science should address the"where" (spatial) and "when" (temporal) questions.

Data are fundamental to the advancement of science, and data sets indexed by space (where) and time (when) allow us to climb higher up the knowledge pyramid than if one or both of these indices were missing. Purely spatial data that do not have a temporal dimension can occur for example when data come from a "snapshot" in time (e.g., liver-cancer rates in USA counties in 2019), or they are taken from a process that is not evolving in time (e.g., an iron-ore body in the Pilbara region of Australia). Sometimes the temporal component has simply been discarded and, at worst, the same may have happened to the spatial component as well.

Purely temporal data sets are not unusual either. For example, two time series of monthly average carbon dioxide (CO_2) measurements, one from the Mauna Loa Observatory, Hawaii, and the other from a global average of CO_2, do not have a spatial dimension (for different reasons). Like the temporal component, sometimes the spatial component is discarded with a consequent loss of information.

Einsteinian Physics

Einstein's theory of relativity (e.g., Bergmann, 1976) demonstrated that space and time are inter-dependent and inseparable. Although we are exclusively concerned with phenomena that reside in a classical Newtonian framework in this paper, we include a brief discussion of space and time within Einstein's framework, to indicate that modifications would be needed for spatio-temporal astronomical data, say.

Einstein proposed the following "thought experiment," a version of which we now give. Think of a boxcar being pulled by a train travelling at velocity v, and place a source of light at the center of the moving boxcar. An observer on the train sees twin pulses of light arrive at the front and rear end of the boxcar *simultaneously.* A stationary observer standing by the train tracks sees one pulse arrive at the rear end of the boxcar *before* its twin arrives at the front end. That is, the reference frame of the observer is extremely important to notions of simultaneity/before/after. What ties together space and time is movement (velocity) of the boxcar.

Einstenian physics assumes that the velocity of light, c, is a universal constant ($\simeq 3 \times 10^5$km/sec), regardless of the spatial frame of reference. Suppose the stationary frame is D^0, and let $\boldsymbol{s}^0 = (x^0, y^0, z^0)'$ represent the Cartesian spatial co-ordinates in D^0. Likewise, D^v is the frame moving at velocity $\boldsymbol{v}$ with Cartesian spatial co-ordinates, $\boldsymbol{s}^v = (x^v, y^v, z^v)'$, where $v \equiv \|\mathbf{v}\|$. Now include the time co-ordinates t^0 and t^v, respectively, and suppose that at $t^0 = t^v = 0$, the two frames have a common origin, $(0, 0, 0)$. A pulse of light sent from that common origin reaches (x^0, y^0, z^0) in time t^0, or equivalently reaches (x^v, y^v, z^v) in time t^v. Since c is constant, we have

$$\{(x^0)^2 + (y^0)^2 + (z^0)^2\}^{1/2}/t^0 = c = \{(x^v)^2 + (y^v)^2 + (z^v)^2\}^{1/2}/t^v \,.$$

That is, no matter which frame of reference is used, the following relationship *always* holds:

$$(x^v)^2 + (y^v)^2 + (z^v)^2 = c^2(t^v)^2 \,; \qquad v \geq 0 \,,$$

which includes the case $v = 0$ (the stationary frame).

Therefore, while Newtonian physics involves objects tracing out paths in four dimensions (three-dimensional space × one-dimensional time), the physics of Einstein shows that space and time are inextricably linked. Other physical properties are modified too. The length of an object as measured in the moving frame is always smaller than or equal to the length of the object measured in the stationary frame, by a factor of $\{1 - (v/c)^2\}^{1/2}$. A similar factor shortens a time interval in a moving frame, leading to the famous conclusion that the crew of a spaceship flying near the speed of light would return in a few (of *their*) years to find their generation had become old.

Einstein's theory of relativity is most certainly important for some phenomena, but in our studies of Earth's environment, we stay within scales of space and time where the physical laws of Newton hold, and we work with a co-ordinate system that is a Cartesian product of space and time.

Spatio-Temporal Data

From a statistical perspective, data that are nearby in space and/or time, are generally positively correlated. In the case of "competition," the opposite may happen (e.g., under big trees only little trees can grow), but the general conclusion is nevertheless that spatio-temporal data should not be modeled as being statistically independent. Tobler (1970) called this notion "the first law of Geography," and presented it as follows: ". . . everything is related to everything else, but near things are more related than distant things." Consequently, an assumption that spatio-temporal data follow the "independent and identically distributed" (iid) statistical paradigm should typically be avoided. Time series models and spatial process models incorporate statistical dependency into the way they capture variability in *temporal data* and *spatial data*, respectively.

Change-of-support

The global/regional/local scales of spatial variability lead to a phenomenon we shall call *change-of-support.* In other disciplines it is known as downscaling/upscaling, or the ecological effect, or the modifiable areal unit problem. It is in fact a manifestation of Simpson's paradox (Simpson, 1951) in a spatial setting. Simpson's paradox, which has a perfectly rational probabilistic explanation, essentially says the following: In a two-way cross-tabulation, the variables (A and B, say) can exhibit a positive statistical dependence, yet when a third variable (C, say) enters and expands the data to a three-way cross-tabulation, the statistical dependence between A and B can be negative for *each* value of C!

In a temporal setting, a causal statistical model built at a three-monthly scale, may have little or no relevance to the mechanisms in play at the daily scale. Day trading on stock markets, based on economic relationships estimated from quarterly trade figures, would probably lead to financial ruin. In a spatial setting, regional climate data may warn, correctly, of a future drought in the Northwest USA (states of Washington and Oregon). However, local orographic effects may favor certain parts of Oregon to the point where above-average rainfall is consistently received there.

Spatial aggregation is ubiquitous: Federal decisions (e.g., carbon "cap and trade") are made at a national scale, state decisions (e.g., California's clean-air regulations) are made at a regional scale, and city-wide decisions (e.g., Tucson's water-conservation policy) are made at a local scale. These decisions are based on data that come from a variety of spatial scales, however an inappropriate statistical analysis that does not respect the change-of-support effect could lead to the adoption of inappropriate policies at the scale that really matters, namely the management and mitigation scale. As we have mentioned, aggregations over time are equally subject to the change-of-support effect, but there has been less discussion of it in the time series literature, perhaps because time series are often already at the scale needed to answer the questions of interest. In that literature, it is referred to as mixed-frequency analysis (Ghysels, 2016).

Objects in a Dynamical Spatial Environment

There are two major ways to view, and hence to model, the evolving spatial environment in which we live. The *object-view* of the world sees individual objects located in a spatial domain and interacting through time with each other, often as a function of their distance apart. Thus, a household and its characteristics make up a unit of interest to census enumerators. This micro-datum is typically unavailable to social scientists, for confidentiality reasons. Consequently, the census data that are released are typically the *number of objects* in small areas, but not the objects' precise locations. That is, a set of count data from small areas is released, which is simply an aggregated version of the object-view of the world. The geographical extent (i.e., spatial support) of a small area can be stored in a Geographical Information System (GIS) as a polygon, and hence the spatial relationships between small areas and their associated counts are preserved in a GIS. (A GIS is a suite of hardware and software tools that feature linked georeferencing in its database management and in its visualization; e.g., Burrough and McDonnell, 1998.) Alternatively, the *field-view* of the world loses sight of the individual objects and potentially has a (multivariate) datum at every spatial location in the domain of interest. Building on the census-enumeration example discussed above, we can define a *field* as the object density, in units of number per unit area, at any location. This is purely a mathematical construct because, at a given location, either there is an object present or there is not. Such a density can be estimated from a moving window, such that at any location the estimated density is proportional to the number of objects per unit area in the window at that location.

Sometimes the field-view is the result of an aggregation of the object-view, such as for population-density data. Other times, the field-view is the only view that is important, such as for rainfall data where there is often no interest in the individual raindrops. Again, a GIS is a convenient way to store data for a field, along with the spatial support to which a datum refers. In general, spatio-temporal data may consist of measurements of both the field type and the object type. Modeling these data with coherent, spatio-temporal, random processes is one of the next frontiers of spatio-temporal statistics (Micheas, 2020).

Uncertainty and the Role of Statistics

Uncertainty is everywhere. As Benjamin Franklin famously said (Sparks, 1840), "In this world nothing can be said to be certain, except death and taxes." Not only is our world uncertain, our attempts to explain the world (i.e., Science) are uncertain. And our measurements of our (uncertain) world are uncertain. Statistics is the "Science of Uncertainty," and it offers a coherent approach to handling the sources of uncertainty referred to above. Here, we use the term *Statistical Science* interchangeably with *Statistics* (with a "capital" S); we use *statistics* (with a "small" s) to refer to summaries of the data.

In this paper, we shall express uncertainty through measures of variability (e.g., variance). Just as the physical and biological sciences have the notions of mass balance and energy balance, Statistical Science has a notion of variability balance. The total variability consists of variability due to *measurement*, variability due to using a (more-or-less uncertain) scientific *model* of how the world works, and variability due to uncertainty on *parameters* that control the variability of the measurements and the models.

Although real-world systems may in principle be partially deterministic, our information is incomplete at each of the stages of observation, summarization, and inference, and thus our understanding is clouded by uncertainty. Consequently, by the time the inference stage is reached, the lack of certainty will influence how much knowledge we can gain from the data. Furthermore, if the dynamics of the system are nonlinear, the processes can exhibit chaos, even though the theory is based on deterministic dynamical systems.

There is a general approach to accounting for uncertainties, and that is through the use of conditional probability models in a hierarchical statistical model. In the next section, we consider uncertainty in the process, as distinct from uncertainty in the measurements of that process. To illustrate the spatio-temporal context, we choose the field-view and a process model governed by a partial differential equation in one-dimensional space that captures the dynamical evolution of the process $\{Y(s;t) : s \geq 0, t \geq 0\}$.

Interaction of space and time

When there is a physical mechanism to be modeled, a partial differential equation (PDE) can be a useful way to describe what happens to infinitesimal changes over time as a function of infinitesimal spatial diffusion. Knowledge of the physics is translated into the form of the PDE. To illustrate this applied-mathematics approach, we consider a reaction-diffusion PDE in one-dimensional space:

$$\frac{\partial Y(s;t)}{\partial t} = \beta \frac{\partial^2 Y(s;t)}{\partial s^2} - \alpha Y(s;t)\,, \tag{1}$$

where β and α are constant (i.e., the PDE is homogeneous).

Equation (1) is deterministic and can be visualized for several different values of α and β; see Figure 1, where the spatial domain has been discretized to $\{0, \ldots, 40\}$, the temporal domain has been discretized to $\{0, \ldots, 80\}$, and an initial condition of $Y(s;0) = I(15 < s < 24)$ is specified.

Equation (1) can be written succinctly as

$$\frac{\partial Y}{\partial t} - \beta \frac{\partial^2 Y}{\partial s^2} + \alpha Y = 0\,,$$

from which a stochastic version can be obtained:

$$\frac{\partial Y}{\partial t} - \beta \frac{\partial^2 Y}{\partial s^2} + \alpha Y = \delta\,, \tag{2}$$

where $\{\delta(s;t) : s \geq 0, t \geq 0\}$ is a mean-zero Gaussian stochastic process that represents model error. Whittle (1986) has called the equation (2) that defines the stochastic

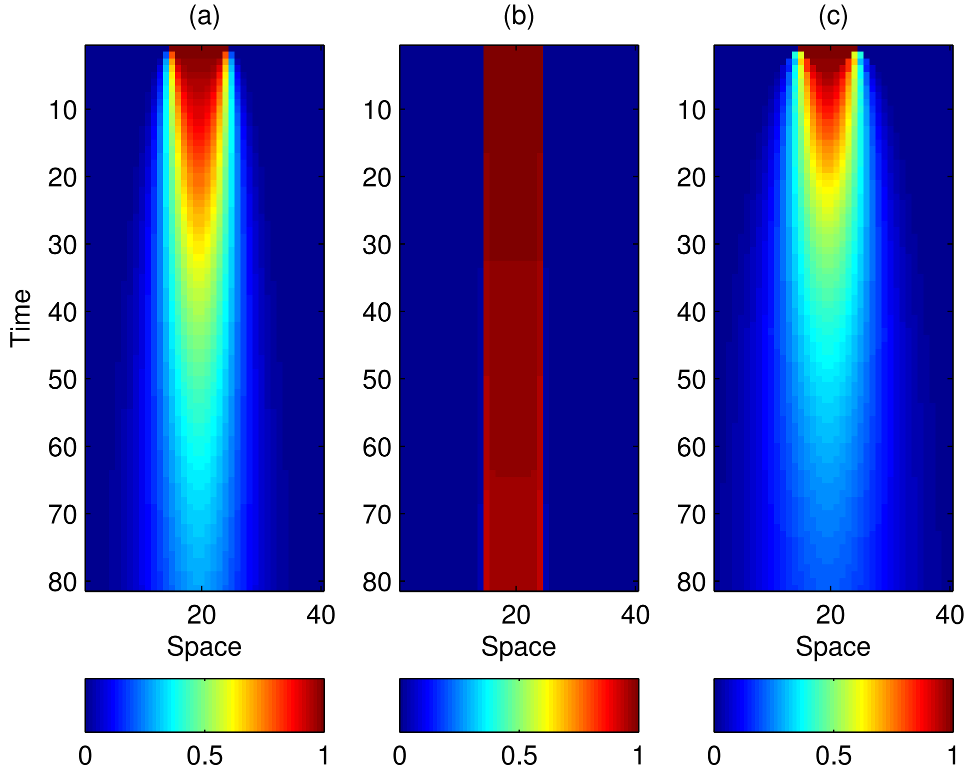

Figure 1: Numerical solution of the reaction-diffusion PDE (1). (a) $\alpha = 1, \beta = 20$; (b) $\alpha = 0.05, \beta = 0.05$; (c) $\alpha = 1, \beta = 50$

process $\{Y(s;t) : s \geq 0, t \geq 0\}$ a diffusion-injection equation. To emphasize the different nature of (2) from (1), we choose the simplest possible process for δ, namely Gaussian white noise: $E(\delta(s;t)) = 0$ and $\text{cov}(\delta(s;t), \delta(u;r)) = \sigma^2 I(s = u, t = r)$, where the parameter σ controls the amount of variability in the stochastic PDE. This specification results in a mean-zero covariance-stationary Gaussian stochastic process Y:

$$C(h;\tau) \equiv \text{cov}(Y(s;t), Y(s+h;t+\tau)\,,$$

with correlation function $\rho(h;\tau) \equiv C(h;\tau)/C(0;0)$; Heine (1955) gives an analytic solution for ρ. Figure 2 gives a single realization of Y for $\alpha = 1$, $\beta = 20$, and three different choices of σ; the smallest value of $\sigma = 0.01$ gives a stochastic realization in Figure 2(a) that looks the most like the deterministic realization in Figure 1(a) where $\sigma = 0$.

The behavior of $Y(s;t)$ for large t illustrates its stationary behaviour. Figure 3 shows three realizations from the Gaussian process with mean zero and $\alpha = 0.05$, $\beta = 20$, $\sigma = 1$ substituted into the covariance function $C(h;\tau)$ corresponding to (2). Simply inspecting a single realization can be misleading if one wishes to infer the parameters α, β, and σ, and we recommend obtaining them from the estimate $\hat{C}(h;\tau)$ of $C(h;\tau)$.

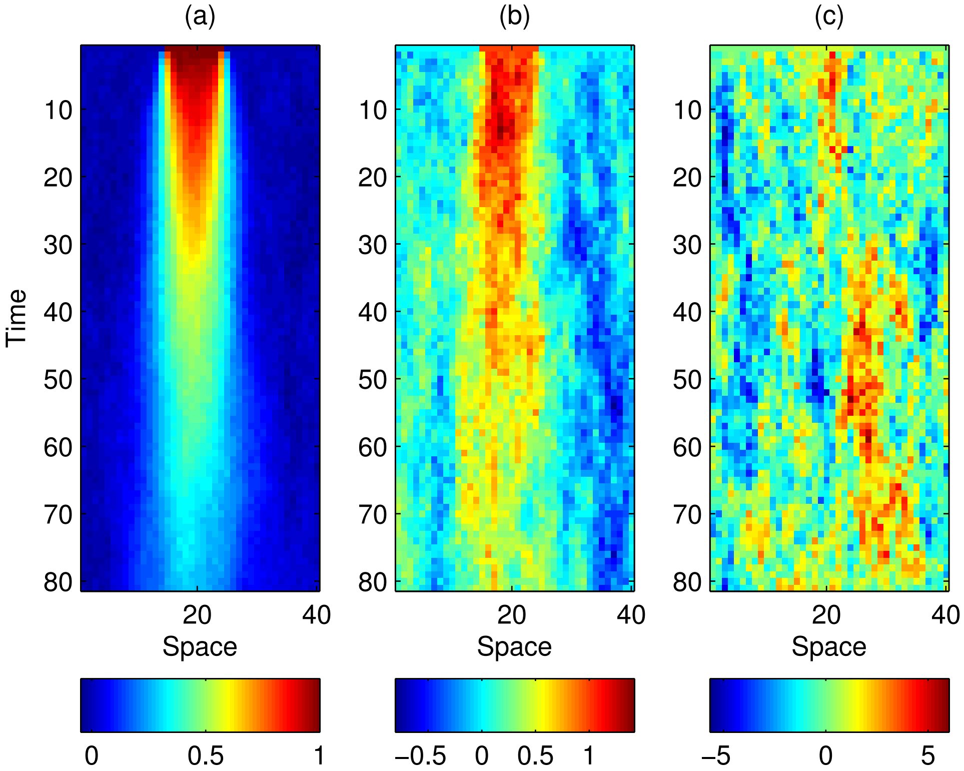

Figure 2: Numerical solution of (2) for one realization of δ and the same initial condition used in Figure 1. (a) $\sigma = 0.01$; (b) $\sigma = 0.1$; (c) $\sigma = 1$, for $\alpha = 1, \beta = 20$.

Now approximate the differentials in the stochastic PDE (2) with differences:

$$\frac{Y(s;t+\Delta_t) - Y(s;t)}{\Delta_t} = \beta\left\{\frac{Y(s+\Delta_s;t) - 2Y(s;t) + Y(s-\Delta_s;t)}{\Delta_s^2}\right\} - \alpha Y(s;t) + \delta(s,t+\Delta_t)\,.$$

Define $\mathbf{Y}_t \equiv (Y(\Delta_s;t),\ldots,Y(39\Delta_s;t)'$ and $\mathbf{Y}_t^B \equiv (Y(0;t),Y(40\Delta_s;t))'$. Then the stochastic difference equation is:

$$\mathbf{Y}_{t+\Delta_t} = \mathbf{M}\mathbf{Y}_t + \mathbf{M}_B\mathbf{Y}_t^B + \boldsymbol{\delta}_{t+\Delta_t}\,,$$

where $\mathbf{M}_B\mathbf{Y}_t^B$ represents given boundary effects, and $\mathbf{M}$ is defined below. For the difference-equation approximation to be stable, bounds on Δ_t and Δ_s are needed, namely $\alpha\Delta_t < 1$ and $2\beta\Delta_t/\Delta_s^2 < 1$. In what follows, we chose $\Delta_t = 0.01$, $\Delta_s = 1$, for which the stability conditions are satisfied when $\alpha = 1$ and $\beta = 20$.

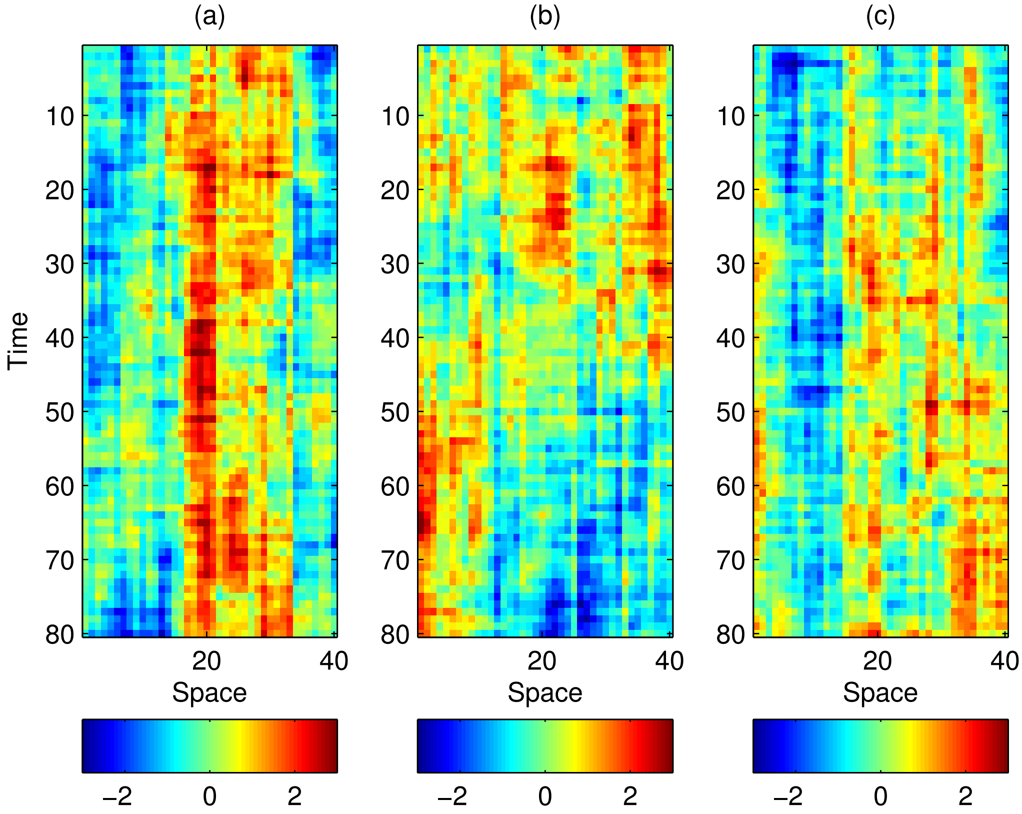

Figure 3: Three realizations generated from a stationary Gaussian stochastic process with covariance function $C(h;\tau)$ for $\alpha = 1$, $\beta = 20$, $\sigma = 1$.

Importantly, the propagator matrix $\mathbf{M}$ has tri-diagonal structure:

$$\mathbf{M} = \begin{bmatrix} \theta_1 & \theta_2 & 0 & \dots & 0 \\ \theta_2 & \theta_1 & \theta_2 & \dots & \vdots \\ 0 & \theta_2 & \theta_1 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \theta_2 \\ 0 & 0 & \dots & \theta_2 & \theta_1 \end{bmatrix},$$

where $\theta_1 = (1 - \alpha\Delta_t - 2\beta\Delta_t/\Delta_s^2)$ and $\theta_2 = \beta\Delta_t/\Delta_s^2$. Note that the matrix $\mathbf{M}$ is defined by the stochastic PDE dynamics, and it is sparse. Conditional on the boundary effects, the temporally lagged covariances are given by,

$$\mathbf{C}_Y^{(m)} = \mathbf{M}^m \mathbf{C}_Y^{(0)}, \tag{3}$$

where $\mathbf{C}_Y^{(m)} \equiv \mathrm{cov}(\mathbf{Y}_t, \mathbf{Y}_{t+m\Delta_t})$, for $m = 0, 1, \ldots$. Figure 4 shows how close the finite-difference approximations given by (3) are to $\rho(h;\tau)$, the analytic solution, for $\alpha = 0.05$, $\beta = 20$, $\sigma = 1$.

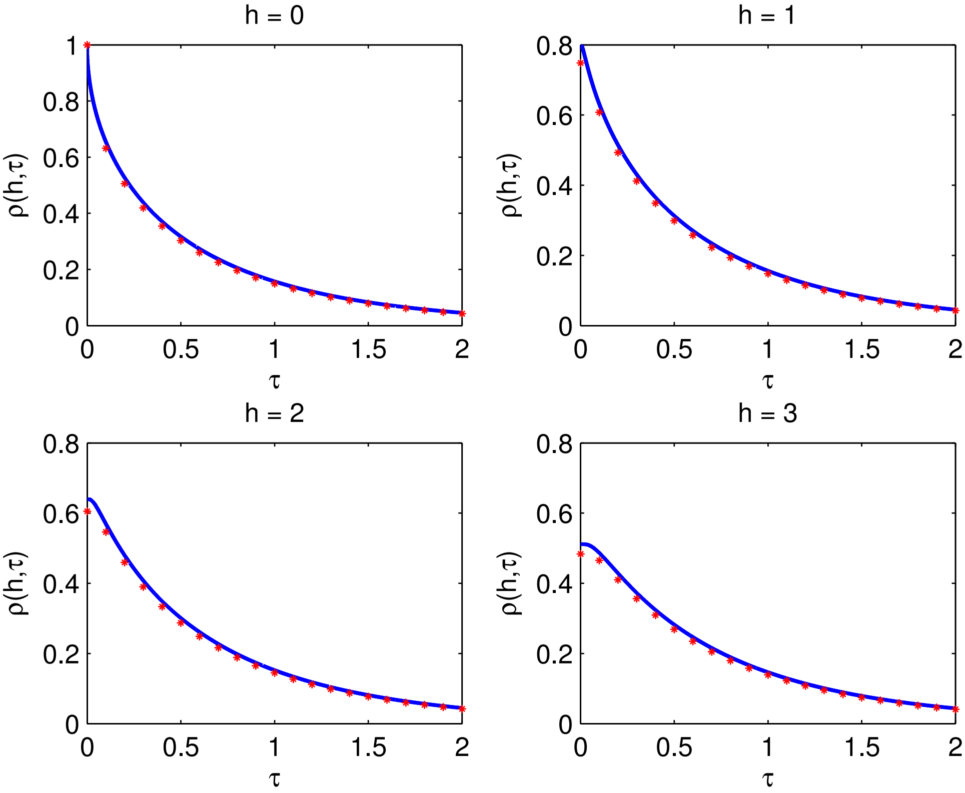

Figure 4: Slices of the spatio-temporal correlations for the stochastic PDE (2) (solid blue line) and the difference approximations (3) (red dots) for $\alpha = 0.05$, $\beta = 20$, and $\sigma = 1$. The four panels show temporal lag τ on the horizontal axis, for the four slices at spatial lags $h = 0, 1, 2, 3$ respectively.

A physically motivated PDE is not always available, but we have spatial intuition from Tobler's first law. That is, we describe the dynamics of $\{\mathbf{Y}_t\}$ according to a vector autoregressive process of order 1 (VAR(1)):

$$\mathbf{Y}_{t+1} = \mathbf{M}\mathbf{Y}_t + \boldsymbol{\delta}_{t+1}\,, \tag{4}$$

where the propagator matrix $\mathbf{M}$ may not depend directly on physically interpretable parameters α and β, but its elements are defined "spatially." That is, m_{ij} are non-zero for s_i and s_j "nearby," and m_{ij} are zero for s_i and s_j "far apart." We shall see below that the VAR(1) model is often a key component of a process model for Y in a spatio-temporal hierarchical model. Now when (4) is inside a hierarchical statistical model, one does not have to solve it, just simulate realizations from it! Consequently, this difference-equation approach, to capturing spatio-temporal dynamics, allows us to model and infer hidden processes of a non-stationary nature, such as

$$Y_{t+1} = \mathbf{M}_{t+1}\mathbf{Y}_t + \boldsymbol{\delta}_{t+1} \,,$$

since simulation is usually quite straightforward. This often represents an advantage over applied-mathematics modeling with continuous time PDEs.

Conditional probabilities in a hierarchical statistical model (HM)

Consider three generic quantities of interest, Z, Y, and θ, which could be random processes, random vectors, or random variables. Think of Z as data, Y as a (hidden or latent) process that we wish to predict, and θ as unknown parameters. In a realistic example where Z, Y, and θ are more complicated random quantities, say for spatial-statistical mapping of a large city's air quality in a given week, Z might be 100-dimensional, Y might be 1000-dimensional, and θ might be five-dimensional. Based on the data Z, we wish to make inference on the latent process Y and on the parameter θ. In a Bayesian hierarchical model (BHM), we wish to predict both Y and θ; and in an Empirical hierarchical model (EHM), we wish to predict Y and to estimate θ. The distribution of Y given the data Z is called the *predictive distribution*.

Bayesian Hierarchical Modeling (BHM)

The basic representation of a BHM is obtained by splitting up the model into three levels (Berliner, 1996):

Data model:	$[Z\vert Y,\theta]$
Process model:	$[Y\vert\theta]$
Parameter model:	$[\theta]$,

where, using generic random quantities A and B, $[A]$ denotes the marginal distribution of A, $[A,B]$ denotes the joint distribution of A and B, and $[A|B]$ denotes the conditional distribution of A given B. Now the joint distribution can be decomposed recursively. From the equation $[A,B]=[A|B][B]$, we have

$$\begin{aligned}[Z,Y,\theta] &= [Z,Y|\theta][\theta] \\ &= [Z|Y,\theta][Y|\theta][\theta], \qquad (5)\end{aligned}$$

which is simply a product of the data model, the process model, and the parameter model.

Bayes' Theorem gives the conditional distribution of Y and θ, given the data Z, which is typically called the *posterior distribution*. The generic result is $[B|A]=[A|B][B]/[A]$, and hence

$$
\begin{aligned}
[Y, \theta | Z] &= \frac{[Z|Y, \theta][Y, \theta]}{\int\int [Z|Y, \theta][Y, \theta] dY d\theta} \\
&= \frac{[Z|Y, \theta][Y|\theta][\theta]}{\int\int [Z|Y, \theta][Y|\theta][\theta] dY d\theta} \\
&= \frac{[Z|Y, \theta][Y|\theta][\theta]}{[Z]}, \qquad (6)
\end{aligned}
$$

where the numerator is precisely (5). Within the framework of Bayesian statistics and decision theory, all inferences on Y and θ in the BHM depends on this posterior distribution.

As we have noted, the numerator in (6) is a product of the individual components of the BHM, but a major problem usually arises when calculating the denominator (or the normalizing constant) that ensures the posterior distribution has total probability equal to 1. (Because the posterior distribution is conditional on Z, in fact the normalizing "constant" depends on the data Z.)

When Y and θ are each single random variables, the integral in the denominator of (6) becomes,

$$
[Z] = \int\int [Z|Y, \theta][Y|\theta][\theta] dY d\theta\,,
$$

which is only a two-dimensional integral and usually quite easy to calculate using numerical quadrature. However, spatio-temporal BHMs can often yield integrals that are of dimensions on the order of thousands (e.g., Wikle et al., 1998). In the last 25 years, computational breakthroughs have been made so that rather than calculating the posterior distribution analytically or numerically, one can often simulate from it. These computational methods, including Markov chain Monte Carlo (MCMC), Hamiltonian Monte Carlo (HMC), Approximate Bayesian Computation (ABC), Variational Bayes (VB), and importance sampling (IS), have brought HM into the panoply of many statisticians, and in particular they have revolutionised the statistical modeling of spatio-temporal data.

Empirical Hierarchical Modeling (EHM)

The following two-level model also qualifies to be called an HM:

Data model:	$[Z\|Y, \theta]$
Process model:	$[Y\|\theta]$,

where it is assumed that the parameter θ is *fixed, but unknown.* Formally, one could still consider a third level, but where the parameter model $[\theta]$ concentrates all its probability at the fixed θ, but that does not help with inference on Y (and θ).

In an EHM, all probability distributions are conditional on a specified value of θ, typically an estimate that depends on the data Z. Inference on Y depends on the distribution (sometimes called the *predictive distribution*)

$$
[Y|Z, \theta] = \frac{[Z|Y, \theta][Y|\theta]}{[Z|\theta]}, \qquad (7)
$$

where $[Z|\theta] = \int [Z|Y,\theta][Y|\theta]dY$, and θ is specified using an estimate $\widehat{\theta}$ or even just a guess. The "Empirical" part of the EHM arises from the practice of replacing (7) with $[Y|Z,\widehat{\theta}]$, where $\widehat{\theta}$ depends only on the data Z. It is also possible that θ is estimated from an independent study.

The difference between (6) and (7) is clear, and which one is used as the posterior distribution depends on the type of HM fitted. Notice that the integral in the denominator in (7) is lower-dimensional than that in the denominator in (6), but it could still be of a dimension on the order of thousands. Importantly, (7) does not require explicit specification of a prior distribution for θ, a task that so-called "frequentist" statisticians are reluctant to do. It can also be the case that (7) is faster to compute or sample from than (6). The price of not specifying uncertainty in the parameter θ is that EHM inferences on Y are generally too liberal, since a simple substitution of $\widehat{\theta}$ for θ does not account for the extra variability associated with the estimation of θ (e.g., Carlin and Louis, 2000, Chapter 4).

Where, When, and then Why

The problem of determining a causative relationship in a process model can be expressed in terms of conditional probabilities. If Y_1 is a phenomenon that could directly affect Y_2 through a physical/chemical/biological/economic/etc. mechanism, and $[Y_2|Y_1]$ changes as Y_1 changes, then Y_1 is a candidate to be a cause of Y_2.

However, even the best of theories can miss an important factor (F, say). This omission might damp down the relationship or yield a negative dependence where there should be a positive one (Simpson, 1951).

Consider the simple process, $Y = (Y_1, Y_2, F)$ made up of three random variables. Then the process model can be written as

$$\begin{aligned}[Y|\theta] &= [Y_1, Y_2|F,\theta][F|\theta] \\ &= [Y_2|Y_1,F,\theta][Y_1|F,\theta][F|\theta]\,.\end{aligned}$$

Thus, the question, "Why does Y_2 behave as it does?" can be answered, "Because Y_1 causes that behavior, and the *type* of dependence is governed by the factor F." If we mistakenly focused on $[Y_2|Y_1,\theta]$ instead of $[Y_2|Y_1,F,\theta]$, we may infer incorrect dependencies. In reality F is present in the mixture distribution,

$$[Y_2|Y_1,\theta] = \int [Y_2|Y_1,F,\theta][F|Y_1,\theta]dF\,.$$

When F represents "level of spatial aggregation," these spurious inferences have been called the *ecological fallacy* (Robinson, 1950). From a spatial-statistical perspective, this problem is part of the research area known as *change-of-support* (e.g., Cressie, 1996, and see the discussion given above in the section, "Change-of-support"). More generally, many factors have spatial and temporal variability; hence, space and time can act as a *proxy* for F, should the process model fail to account for it. In other words, modeling spatio-temporal variability, along with good experimental design, is an "insurance policy" against missing an F, and it should get us closer to Science's holy grail of discovering *causation.*

Experimental Data

Earth's population is many billions and the demand for sustenance is great and continuous. The planet's ability to produce food on a massive scale largely came from fundamental experiments in crop science in the early twentieth century. Fisher (1935) developed a statistical theory of experimental design, based on the three principles of *blocking*, *randomization*, and *replication*, which has allowed farmers to grow high-yielding, insect-resistant crops adapted to local conditions. He developed a vocabulary that is used today in scientific experiments of all types: response (e.g., wheat yields), treatments (e.g., varieties of wheat), factors (e.g., soil type, field aspect, growing season), levels of factors (e.g., for the soil-type factor, the levels might be sand, gravel, silt, clay, peat), plot (experimental unit that receives a single treatment), block (collection of plots with the same factor/level combination), randomization (random assignment of treatments to plots), replication (number of responses per treatment), and so forth.

Data from designed experiments, when analyzed appropriately, allow stronger, (almost) causative inferences, which incubate further inspiration and hypothesis generation, and so forth through the scientific-discovery cycle. In the right hands, and with a component of luck, this cycle can lead to great breakthroughs (e.g., the discovery of penicillin in 1928 by Alexander Fleming). Even small breakthroughs are bricks that are laid on the path from *information* to *knowledge.*

Space and time are fundamental factors of any experiment. For example, "soil type" is highly spatial and "growing season" is highly temporal. The protocol for any well designed experiment should involve recording the location and time at which each datum was collected, because so many factors (known or unknown) correlate with them. After the experiment has been performed, spatial and temporal information can be used as proxies for unknown, unaccounted-for factors (that may later become "known" as the experiment proceeds). From this point of view, the natural place to put spatial and temporal effects in the statistical model is in the mean. But, there is another approach...

In R.A. Fisher's path-breaking work on design of experiments in agricultural science, he wrote (Fisher, 1935, pp. 66): "After choosing the area we usually have no guidance beyond the widely verified fact that patches in close proximity are commonly more alike, as judged by the yield of crops, than those which are further apart." Spatial variability, which to Fisher came in the form of plot-to-plot variability, is largely due to physical properties of the soil and environmental properties of the field. Fisher avoided the confounding of treatment-effect with plot-effect by introducing *randomization* into the scientific method. It was a brilliant insertion of more uncertainty into a place in the experiment where uncertainty abounds, leaving the more certain parts of the experiment intact. Fisher's idea has had an enormous effect on all our lives. For example, any medicine we have taken to treat our ailments and illnesses has gone through rigorous testing, to which the *randomized* clinical trial is central (where a "plot" is often the patient).

Randomization comes with a price. It allows valid inference on the treatments through a simple expression for the mean response, but the variances and covariances of the responses are affected too. Under randomization of the assignment of treatments to plots, the notions of "close proximity" and "far apart" have been hustled out the back door. Can we get spatial dependence back into the statistical analysis of responses, resulting in more efficient inferences for treatment effects? Papadakis (1937) and Bartlett (1938) gave spatial analyses for data from agricultural field trials, and Fairfield Smith (1938) formulated the problem of choosing an optimal plot size based on empirical observations that could have only made sense in the presence of spatial correlation. But spatial statistical models for such phenomena did not begin to appear until much later (e.g., Whittle, 1954; Besag, 1974; McCullagh and Clifford, 2006). It has become clear in the last two decades that spatial statistical analyses do not usurp the standard designs and analyses of variance proposed by Fisher, but they can augment them in search of more efficient inferences and hence shorter, less costly experiments (e.g., Grondona and Cressie, 1991, 1993; Brownie and Gumpertz, 1997; Federer et al., 1997; Legendre et al., 2004).

Observational Data

Organisms are born, live, reproduce, and die but, in the process, they can produce harmful by-products that may threaten their own well being as well as the well being of other organisms around them. *Ecology* is the study of these organisms, and *environment* is their surroundings. Variability within organisms can be large, as can variability between their environments. Thus, it can be very difficult to conduct controlled experiments on Earth's ecology and environment.

Observational data come from a "wilder side" of Science. The environment (such as climate, air and water quality, radioactive contamination, etc.) is a part of our lives that often will not submit to blocking, randomization, and replication. We cannot control when it rains, nor can we observe two Los Angeles, one with smog and one without. However, we can look for two like communities, one with contaminated water and one without; and we can look at health records before and after a toxic emission. Nevertheless, any inference is tentative because the two factors, space and time, are not controlled for. Collecting samples from ambient air presents a philosophical problem because the parcel of air is unique when it passes the monitoring site; it evolves as the changes in air pressure move it around, and it will never come back to allow us the luxury of obtaining an independent identically distributed observation.

In the environmental and life sciences, classical experimental design can struggle to keep up with answering the questions being asked, but they still need to be answered. And, as we have discussed just above, uncertainty is likely to be higher without experimental control. Thus, Statistics has a crucial role to play, although for observational data it does not fit neatly into the blocking-randomization-replication framework. Even when one is able to "block" human subjects on age and sex, say, it may be that an unknown genetic factor will determine how a patient responds to a given treatment. In epidemiological studies, controls may be randomly matched with cases, but the cases are in no way assigned randomly to neighborhoods. And, although duplicate chemical assays allow for assessment of measurement error in a study on stream pollution, replication of a water parcel of the same volume from the stream is impossible. In circumstances such as these, Statistics is even more relevant.

In the environmental sciences, proximity in space is a particularly relevant factor. The word "environ" means "around" in French. "Proximity" is a relative notion, relative to the spatial scale of the phenomenon under study. A toxic-waste disposal site may directly affect a neighborhood of a few square miles; a coal-burning power plant may directly affect a heavily populated region of hundreds of square miles, and an increase in greenhouse gases will affect the whole planet.

Clearly, a *global* effect is felt locally in many ways, and a quantity like global mean temperature is a largely uninformative summary of how daily lives of a community will be affected by a warmer planet, which means that environmental studies of the globe must recognize the importance of *local* variability. Further, how the spatial variability behaves dynamically (i.e., the spatio-temporal variability) is key to understanding the causes of global warming and what to do about it. Finally, we state the obvious, that political boundaries cannot hold back a one-meter rise in sea level; our environment is ultimately a global resource and its stewardship is an international responsibility.

Dynamic HMs and Connections to Deep Neural Models

One of the fundamental principles of hierarchical dynamic spatio-temporal models (H-DSTMs), starting with Wikle et al. (1998), is that to model complex processes across multiple time and spatial scales we must consider a multi-level sequence of linked conditional probability models (see Cressie and Wikle, 2011, for an overview). In particular, because it is very difficult to specify the dependence structure for complex processes directly, one builds it by placing the modeling effort into the conditional mean and capturing statistical dependence through marginalization. Similarly, the deep neural network models in machine learning that have become popular in the last decade for image and language processing (e.g., convolutional neural networks (CNNs) and recurrent neural networks (RNNs)) are also based on a sequence of linked models (although typically not stochastic), with the outputs from one level becoming the inputs for the next (e.g., see Goodfellow et al., 2016, for an overview). Spatio-temporal versions of these models, what might be called deep neural DSTMs (DN-DSTMs), typically combine CNNs and RNNs and seek to build complexity by learning which scales of spatial and/or temporal variability are important for predicting responses (e.g., Donahue et al., 2015).

As discussed in Wikle (2019), the machine-learning DN-DSTM framework has many features in common with the statistical H-DSTM framework. For example, both model frameworks typically (a) consist of multiple connected levels; (b) include dimension reduction; (c) do not model second-order dependence directly; (d) can handle multiple inputs and different output types; (e) require a very large number of parameters to be estimated; (f) require an abundance of training data; (g) require prior information (or, pre-training, heuristics, etc.); (h) require regularization; and (i) are expensive to compute and require efficient algorithmic implementations. Although there are these common features between the modeling paradigms, there are some fundamental differences as well. For example, the H-DSTM is constructed with stochastic models that include distributional error terms within a coherent probability framework (i.e., the joint distribution of all random components can be written as a product of a series of conditional models). In contrast, the DN-DSTM is typically deterministic with no error terms. A consequence of this lack of probabilistic structure is that there is no clear mechanism to produce model-based estimates of uncertainty in the prediction or classification that results from using a DN-DSTM. In addition, one is not able to perform inference on the parameters in a DN-DSTM, although this has not been of particular interest since prediction/attribution is the goal and parameters are typically not identifiable, highly dependent, and non-interpretable. Recently, there have been research initiatives to interpret and explain fitted deep neural networks; see Samek et al. (2019).

Given the similarities and differences discussed above, there has been recent interest in combining the statistical formality of H-DSTMs and the flexibility of DN-DSTMs into a framework that allows complex processes to be modeled in a fairly parsimonious manner and with computational efficiency. For example, McDermott and Wikle (2017) take a simple ensemble-forecast approach (analogous to a parametric bootstrap) in which they simply consider multiple samples from a special type of RNN known as an Echo State Network (ESN), which is notable for the fact that the parameters that describe the dynamics are not learned but rather are chosen randomly. By including quadratic outputs and borrowing the notion of "embeddings" from dynamical systems (e.g., Takens, 1981), they were able to outperform state-of-the-art H-DSTMs for long-lead prediction of Pacific sea surface temperatures (i.e., El Niño/La Niña), and they were able to do this at a fraction of the computational cost (just seconds to implement on a laptop compared to hours for traditional H-DSTM or DN-DSTM approaches.)

The approach of McDermott and Wikle (2017) has no mechanism to link multiple hidden layers, which is important for processes that occur on multiple time scales. McDermott and Wikle (2019) resolve this by considering ensembles of deep ESN models as basis-function generators (after all, a neural network, even a deep one, is simply a nonlinear transformation of the inputs). They then use these basis functions within a regularized generalized additive mixed model, or within a classical Bayesian hierarchical model with stochastic variable-selection priors. Essentially, this represents a high-dimensional (auto)regression problem in which an abundance of basis functions are generated by stochastic transformation of the inputs through the deep ESN process. Multiple transformations are considered as potential predictors to provide flexibility and reproducibility, and then the number of predictors are controlled by regularization. It is important to note that the inputs (predictors) are stochastically and dynamically transformed so that, even if the spatio-temporal regression model is not itself dynamic, the transformations are dynamic through the ESN structure. These multiple levels of transformations allow for different time and spatial scales in the predictor variables to affect the response. Thus, this approach can model very complex spatio-temporal processes very efficiently.

The use of deep neural models can also be used to facilitate implementation of a wide variety of spatial or spatio-temporal statistical analyses. For example, Zammit-Mangion et al. (2019) consider deep Gaussian processes to efficiently estimate spatial warping functions that transform non-stationary spatial processes into stationary spatial processes. More recently, Zammit-Mangion and Wikle (2020) use convolution neural networks to train a state-dependent transition operator in an integro-difference-equation-based DSTM. They show that this model has the remarkable ability to forecast dynamical systems that are completely different from the one upon which it was trained (so-called "transfer learning"). The blending of deep machine-learning models and statistical models is one of the new frontiers in spatial and spatio-temporal statistics.

Concluding remarks

The first author was honored to give the 2019 Santálo Lecture at Complutense University, Madrid, in October 2019. The last part of his lecture put the philosophy of measuring, mapping, and uncertainty quantification into practice, specifically in the area of remote sensing of the environment from NASA's Orbiting Carbon Observatory-2 satellite. Since 2014, it has been on a mission to estimate the global geographic distribution of carbon dioxide sources and sinks at Earth's surface, through time. A description of that remote-sensing research can be found in Cressie (2018), where geolocated measurements from the satellite were used in an HM to map predicted total-column carbon dioxide over the whole globe. The HM allows the production of a second map that quantifies the uncertainty in the predicted values in the original map.

Acknowledgements

Material from Chapters 1, 2, and 6 of Cressie and Wikle (2011) is used with permission from the publisher: Copyright ©(2011) by John Wiley & Sons, Inc. All rights reserved.

References

Bartlett, M. S. (1938). The approximate recovery of information from replicated field experiments with large blocks. *Journal of Agricultural Science*, 28:418–427.

Bergmann, P. G. (1976). *Introduction to the Theory of Relativity.* Dover, New York, NY.

Berliner, L. M. (1996). Hierarchical Bayesian time-series models. In *Maximum Entropy and Bayesian Methods*, pages 15–22. Kluwer Academic Publishers, Dordrecht, NL.

Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, 36:192–225.

Brownie, C. and Gumpertz, M. L. (1997). Validity of spatial analyses for large field trials. *Journal of Agricultural, Biological, and Environmental Statistics*, 2:1–23.

Burrough, P. and McDonnell, R. A. (1998). *Principles of Geographical Information Systems, 2nd edn.* Oxford University Press, Oxford, UK.

Carlin, B. P. and Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman and Hall/CRC, Boca Raton, FL.

Cressie, N. (1996). Change of support and the modifiable areal unit problem. *Geographical Systems*, 3:159–180.

Cressie, N. (2018). Mission CO$_2$ntrol: A statistical scientist's role in remote sensing of atmospheric carbon dioxide (with discussion). *Journal of the American Statistical Association*, 113:152–181.

Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data.* John Wiley & Sons, Hoboken, NJ.

Donahue, J., Hendricks, A. L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634.

Fairfield Smith, H. (1938). An empirical law describing heterogeneity in the yields of agricultural crops. *Journal of Agricultural Science*, 28:1–23.

Federer, W. T., Newton, E. A., and Altman, N. S. (1997). Combining standard block analyses with spatial analyses under a random effects model. In *Modelling Longitudinal and Spatially Correlated Data*, pages 373–386. Springer, New York, NY.

Fisher, R. A. (1935). *The Design of Experiments.* Oliver and Boyd, Edinburgh, UK.

Ghysels, E. (2016). Macroeconomics and the reality of mixed frequency data. *Journal of Econometrics*, 193:294–314.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning.* MIT Press.

Grondona, M. O. and Cressie, N. (1991). Using spatial considerations in the analysis of experiments. *Technometrics*, 33:381–392.

Grondona, M. O. and Cressie, N. (1993). Efficiency of block designs under stationary second-order autoregressive errors. *Sankhyā A*, 55:267–284.

Heine, M. (1955). Models for two-dimensional stationary stochastic processes. *Biometrika*, 42:170–178.

Legendre, P., Dale, M. R. T., Fortin, M.-J., Casgrain, P., and Gurevitch, J. (2004). Effects of spatial structures on the results of field experiments. *Ecology*, 85:3202–3214.

McCullagh, P. and Clifford, D. (2006). Evidence for conformal invariance of crop yields. *Proceedings of the Royal Society, Series A*, 462:2119–2143.

McDermott, P. L. and Wikle, C. K. (2017). An ensemble quadratic echo state network for non-linear spatio-temporal forecasting. *Stat*, 6:315–330.

McDermott, P. L. and Wikle, C. K. (2019). Deep echo state networks with uncertainty quantification for spatio-temporal forecasting. *Environmetrics*, 30:e2553.

Micheas, A. (2020). *Theory and Modeling of Stochastic Objects: Point Processes and Random Sets*. Chapman & Hall/CRC Press, Boca Raton, forthcoming.

Papadakis, J. (1937). Méthode statistique pour des expériences sur champ. *Bulletin Scientifique No. 23*, pages 13–29, Institut d'Amélioration des Plantes à Salonique, Greece.

Robinson, W. S. (1950). Ecological correlation and the behavior of individuals. *American Sociological Review*, 15:351–357.

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Müller, K.-R. (2019). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer Nature, Heidelberg, Germany.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B*, 13:238–241.

Sparks, J. (1840). *The Works of Benjamin Franklin, Vol. 10*. Hilliard, Gray & Company, Boston, MA.

Takens, F. (1981). Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence, Warwick 1980*, pages 366–381. Springer, Berlin, DE.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46:234–240.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, 41:434–449.

Whittle, P. (1986). *Systems in Stochastic Equilibrium*. John Wiley and Sons, Chichester, UK.

Wikle, C. K. (2019). Comparison of deep neural networks and deep hierarchical models for spatio-temporal data. *Journal of Agricultural, Biological and Environmental Statistics*, 24:175–203.

Wikle, C. K., Berliner, L. M., and Cressie, N. (1998). Hierarchical Bayesian space-time models. *Environmental and Ecological Statistics*, 5:117–154.

Wikle, C. K., Zammit-Mangion, A., and Cressie, N. (2019). *Spatio-Temporal Statistics with R*. Chapman & Hall/CRC Press, Boca Raton, FL.

Zammit-Mangion, A., Ng, T. L. J., Vu, Q., and Filippone, M. (2019). Deep compositional spatial models. *arXiv Preprint (arXiv:1906.02840)*.

Zammit-Mangion, A. and Wikle, C. K. (2020). Deep integro-difference equation models for spatio-temporal forecasting. *Spatial Statistics*, forthcoming.