# Spatio-Temporal Statistics: Brief Introduction and Modern Challenges

Christopher K. Wikle
Curators' Distinguished Professor and Chair of Statistics

University of Missouri

18 July, 2023

# Introduction to Spatio-Temporal Statistics

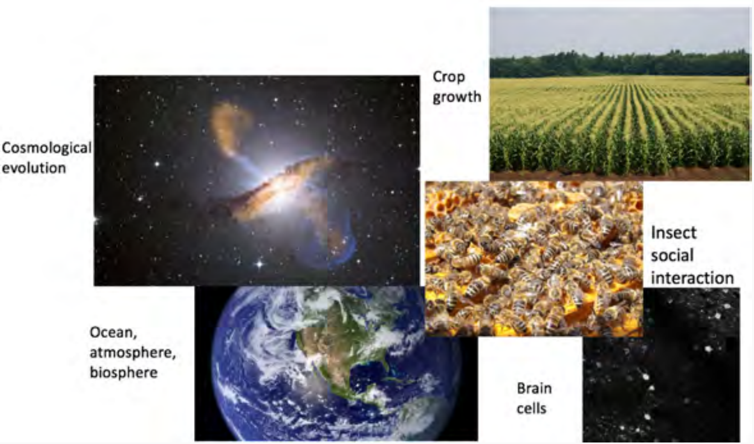
- For a gentle introduction, see "Spatio-temporal Statistics with R" by Christopher K. Wikle, Andrew Zammit-Mangion, and Noel Cressie (2019).
- Book downloadable for free from https://spacetimewithr.org
- Practical R labs for each chapter
- Some of these slides are extracted from a copyrighted short course developed by the authors.

# Talk Outline

- Motivation and goals of spatio-temporal statistics
- Overview of descriptive (covariance-based) spatio-temporal statistics
- Overview of dynamical (conditional) spatio-temporal statistics
- Linear dynamic spatio-temporal models
- Nonlinear dynamic spaito-temporal models
- Hybrid neural/statistical approaches

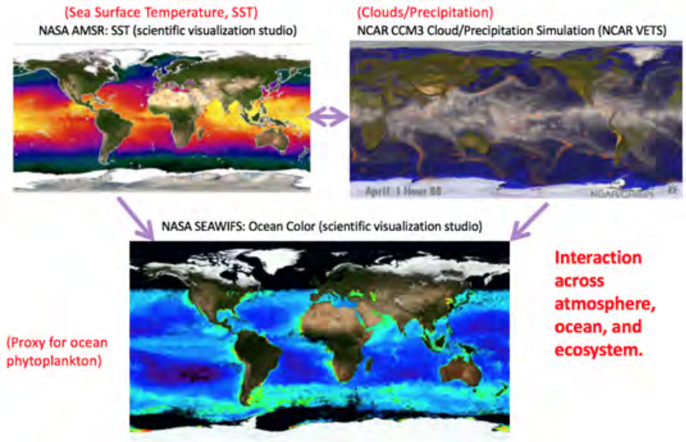
# Complex System

Our universe is a complex system of interacting physical, biological, and social processes across a huge range of time and spatial scales of variability!



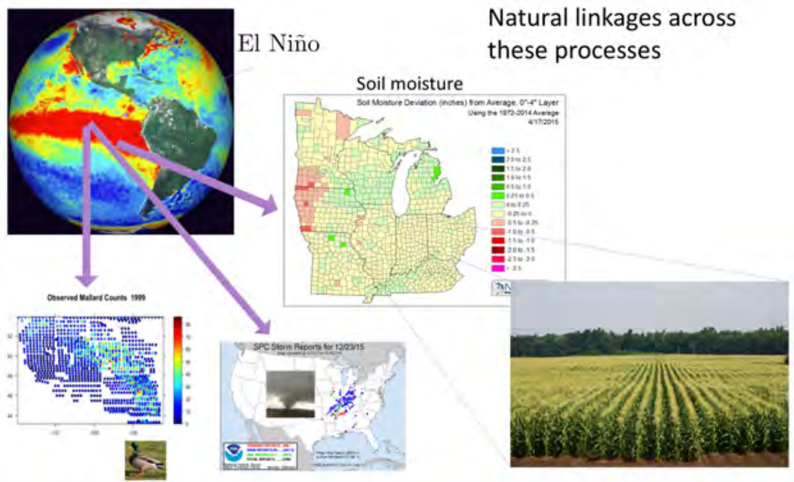
# Interactions Across Space and Time

It is often not sufficient to consider just a snapshot of a spatial process at a given time, or time series at a spatial location or an average over spatial locations.


(Sea Surface Temperature, SST)
NASA AMSR: SST (scientific visualization studio)
(Clouds/Precipitation)
NCAR CCM3 Cloud/Precipitation Simulation (NCAR VETS)
April 1 Hour 00
NASA SEAWIFS: Ocean Color (scientific visualization studio)
(Proxy for ocean phytoplankton)
Interaction across atmosphere, ocean, and ecosystem.

# Interaction and Impacts

Interaction and impacts across processes and spatio-temporal scales.


El Niño
Natural linkages across these processes
Soil moisture
Soil Moisture Deviation (inches) from Average, 0"-4" Layer
Using the 1972-2014 Average
4/17/2015
> 2.5
2.0 to 2.5
1.5 to 2.0
1.0 to 1.5
0.5 to 1.0
0.25 to 0.5
0 to 0.25
-0.25 to 0
-0.5 to -0.25
-1.0 to -0.5
-1.5 to -1.0
-2.0 to -1.5
-2.5 to -2.0
< -2.5
Observed Mallard Counts 1999
SPC Storm Reports for 12/23/15

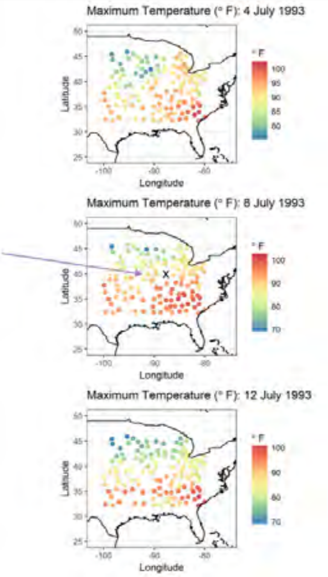
# Goals of Spatio-Temporal Analysis

Characterize spatio-temporal processes in the presence of uncertain and (often) incomplete observations and system knowledge, for the purposes of:

- Prediction in space
- Forecasting in time
- Assimilation of observations and mechanistic models
- Accounting for dependence in parameter inference
- Classification

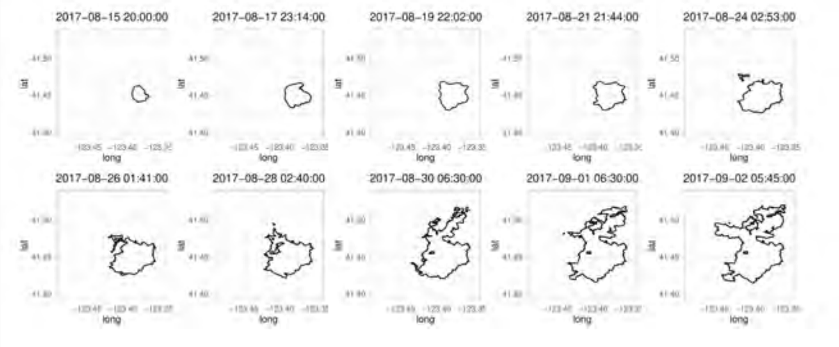
# Example: Prediction at a Spatial Location

What is the optimal prediction (interpolation) to get temperature at the location given by “×”?


Maximum Temperature (° F): 4 July 1993
Maximum Temperature (° F): 8 July 1993
Maximum Temperature (° F): 12 July 1993
Latitude
Longitude
50
45
40
35
30
25
-100
-90
-80
° F
100
95
90
85
80
70
X

# Example: Forecast in Time

Forecasting the spread of a mega wildfire



# Example: Parameter Estimation

What habitat characteristics influence the spread of rabies in the northeast US?

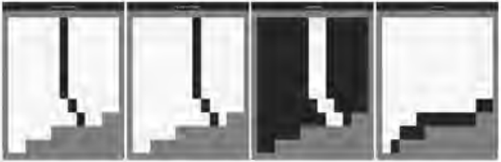
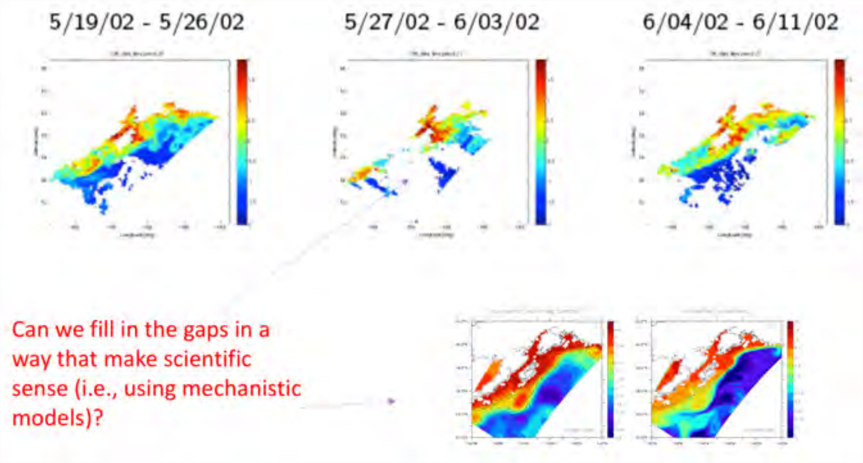


Figure 6: Plots showing the local covariate information from left - counties that neighbor the Connecticut River (west and east sides), non-river counties, and coastal counties. Note that the jagged southern edge is the Long Island Sound and the west, north, and eastern edges are the states of New York, Massachusetts, and Rhode Island, respectively.

# Example: Assimilation of Observations and Mechanistic Models



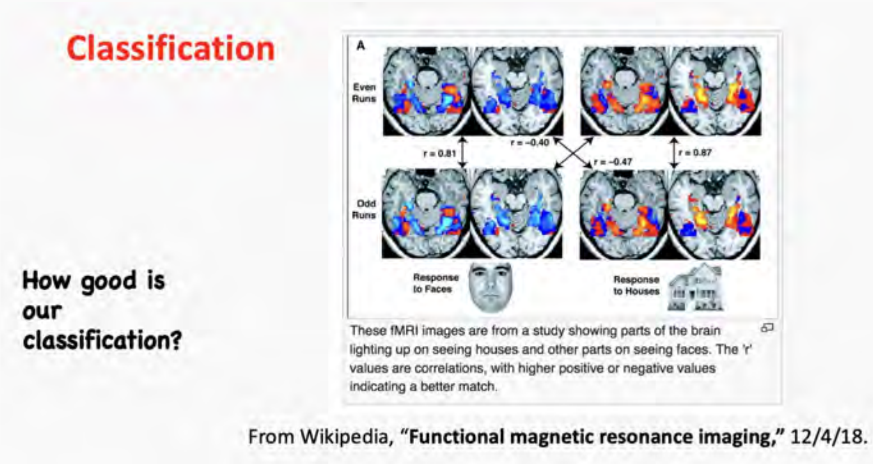
5/19/02 - 5/26/02

5/27/02 - 6/03/02

6/04/02 - 6/11/02

Can we fill in the gaps in a way that make scientific sense (i.e., using mechanistic models)?

# Example: Classification



## Classification

How good is our classification?

These fMRI images are from a study showing parts of the brain lighting up on seeing houses and other parts on seeing faces. The 'r' values are correlations, with higher positive or negative values indicating a better match.

From Wikipedia, **"Functional magnetic resonance imaging,"** 12/4/18.

# Challenges in Spatio-Temporal Modeling

- Accommodating complex dependence (e.g., multiple scales in space and time; mechanistic plausibility; nonlinearity)
- Uncertainty in observations, process knowledge, parameterizations (quantify this uncertainty!)
- Dimensionality
- Missing observations and change of support (resolution)
- Non-Gaussian observations
- Multiple data streams
- Others, e.g., nonlinearity, non-Markovian dependence, etc.

# Spatio-Temporal Modeling

Traditionally, there are two approaches for spatio-temporal modeling to address these goals: the “two D’s”

Descriptive

Dynamic

# Spatio-Temporal Modeling

**Descriptive (Marginal) Approach:**

Characterize the first- and second-moment behavior of the process

- Convenient “optimal” prediction theory
- Precedence in spatial statistics
- Need only specify mean structure and covariability
- Most useful when knowledge of the process is limited and/or primary interest is with inference on fixed-effects parameters
- Can be difficult computationally in high dimensions
- Difficult to specify realistic spatio-temporal covariances for complex processes

# Spatio-Temporal Modeling: Descriptive Approach

**Tobler's (1970) "First Law of Geography"** (spatio-temporal extension):

> *"everything is related to everything else, but near things [in space and time] are more related than distant things"*

(Except when they aren't!)

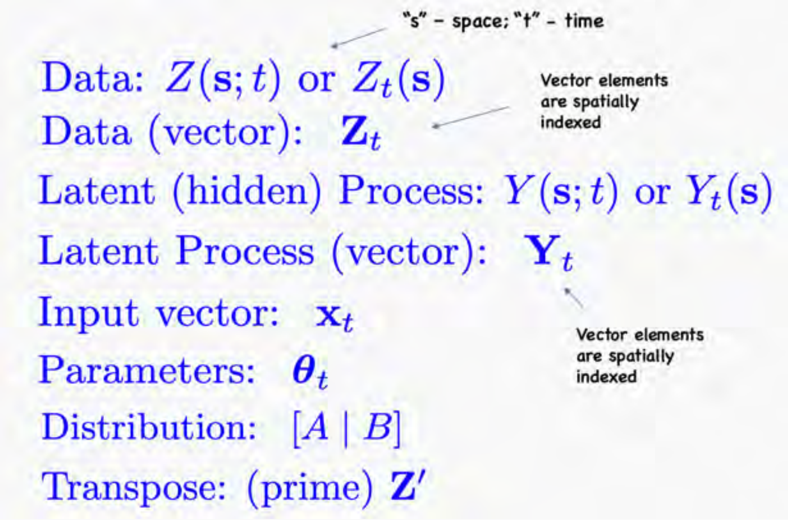
# Spatio-Temporal Modeling: Descriptive Approach

Much of the literature in spatial statistics in the last 10-20 years has been concerned with how to implement descriptive spatial prediction with very large data sets and large prediction domains, as well as the development of new classes of covariance functions.

You have seen examples of this in the course so far.

Let us recap the essence of descriptive modeling, but focus on the spatio-temporal setting.

# Notation



"s" – space; "t" – time

Data: $Z(\mathbf{s};t)$ or $Z_t(\mathbf{s})$

Data (vector): $\mathbf{Z}_t$

Vector elements are spatially indexed

Latent (hidden) Process: $Y(\mathbf{s};t)$ or $Y_t(\mathbf{s})$

Latent Process (vector): $\mathbf{Y}_t$

Vector elements are spatially indexed

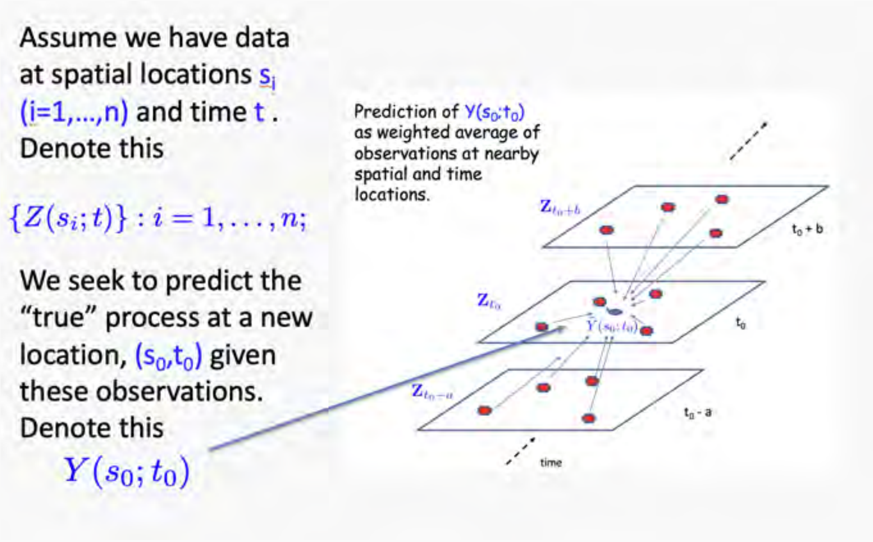
Input vector: $\mathbf{x}_t$

Parameters: $\boldsymbol{\theta}_t$

Distribution: $[A \mid B]$

Transpose: (prime) $\mathbf{Z}'$

# Brief Overview: Descriptive Spatio-Temporal Modeling



Assume we have data at spatial locations $s_i$ $(i=1,\ldots,n)$ and time $t$ . Denote this

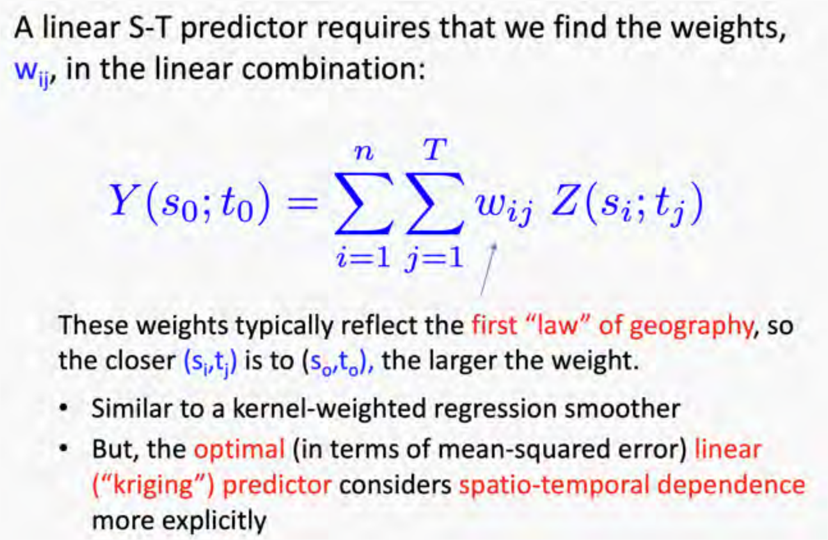
$$\{Z(s_i;t)\} : i = 1, \ldots, n;$$

We seek to predict the “true” process at a new location, $(s_0,t_0)$ given these observations. Denote this

$$Y(s_0;t_0)$$

Prediction of $Y(s_0;t_0)$ as weighted average of observations at nearby spatial and time locations.

# Brief Overview: Descriptive Spatio-Temporal Modeling



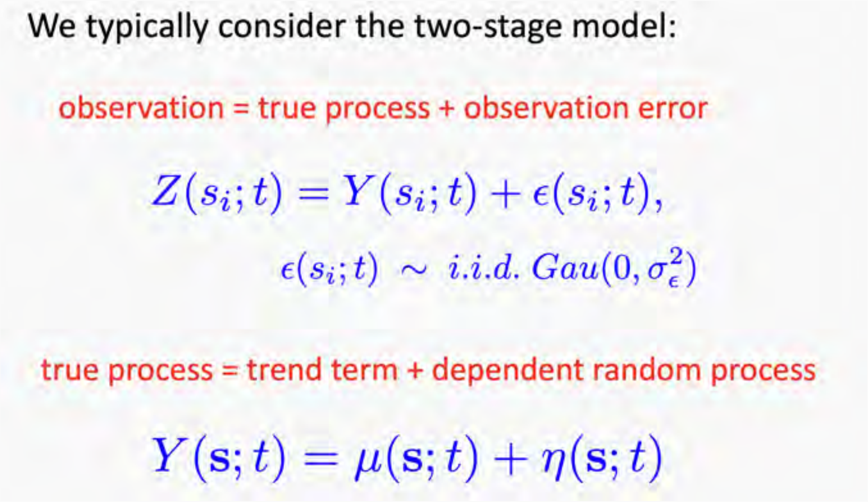
A linear S-T predictor requires that we find the weights, $w_{ij}$, in the linear combination:

$$Y(s_0;t_0) = \sum_{i=1}^{n} \sum_{j=1}^{T} w_{ij}\ Z(s_i;t_j)$$

These weights typically reflect the first "law" of geography, so the closer $(s_i,t_j)$ is to $(s_o,t_o)$, the larger the weight.

- Similar to a kernel-weighted regression smoother
- But, the optimal (in terms of mean-squared error) linear ("kriging") predictor considers spatio-temporal dependence more explicitly

# Brief Overview: Descriptive Spatio-Temporal Modeling



We typically consider the two-stage model:

observation = true process + observation error

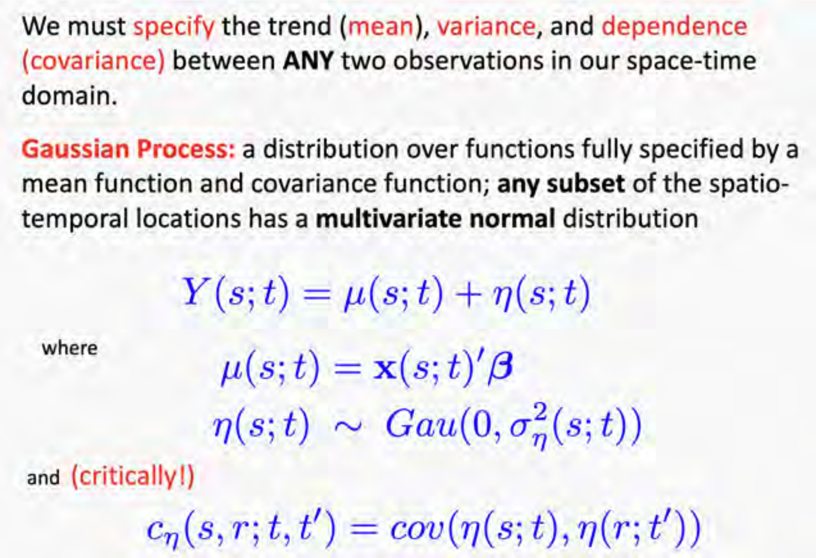
$$Z(s_i;t) = Y(s_i;t) + \epsilon(s_i;t),$$

$$\epsilon(s_i;t) \ \sim \ i.i.d.\ Gau(0, \sigma_\epsilon^2)$$

true process = trend term + dependent random process

$$Y(\mathbf{s};t) = \mu(\mathbf{s};t) + \eta(\mathbf{s};t)$$

# Brief Overview: Descriptive Spatio-Temporal Modeling



We must specify the trend (mean), variance, and dependence (covariance) between **ANY** two observations in our space-time domain.

**Gaussian Process:** a distribution over functions fully specified by a mean function and covariance function; **any subset** of the spatio-temporal locations has a **multivariate normal** distribution

$$Y(s;t) = \mu(s;t) + \eta(s;t)$$

where

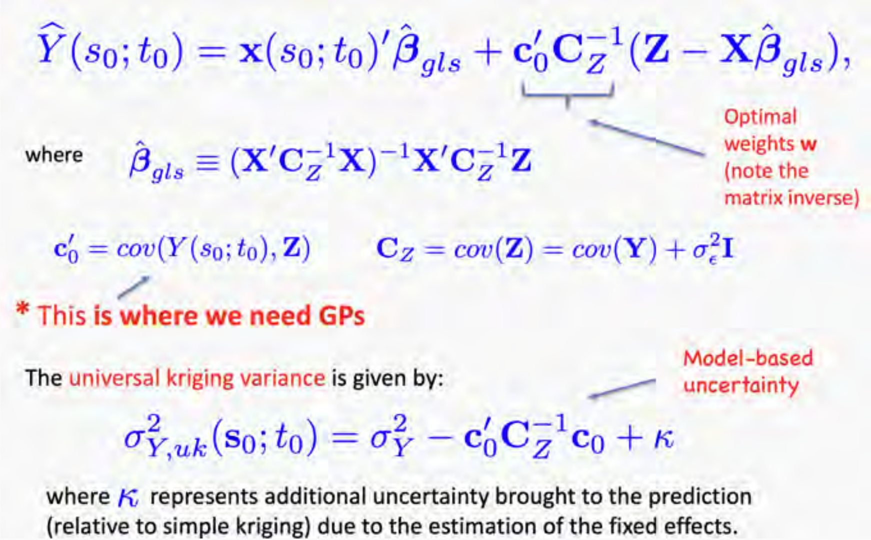
$$\mu(s;t) = \mathbf{x}(s;t)'\boldsymbol{\beta}$$

$$\eta(s;t) \;\sim\; Gau(0, \sigma_\eta^2(s;t))$$

and (critically!)

$$c_\eta(s,r;t,t') = cov(\eta(s;t), \eta(r;t'))$$

# Brief Overview: Descriptive Spatio-Temporal Modeling



$$\widehat{Y}(s_0;t_0) = \mathbf{x}(s_0;t_0)'\hat{\boldsymbol{\beta}}_{gls} + \underbrace{\mathbf{c}_0'\mathbf{C}_Z^{-1}}(\mathbf{Z} - \mathbf{X}\hat{\boldsymbol{\beta}}_{gls}),$$

Optimal weights **w** (note the matrix inverse)

where $$\hat{\boldsymbol{\beta}}_{gls} \equiv (\mathbf{X}'\mathbf{C}_Z^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}_Z^{-1}\mathbf{Z}$$

$$\mathbf{c}_0' = cov(Y(s_0;t_0),\mathbf{Z}) \qquad \mathbf{C}_Z = cov(\mathbf{Z}) = cov(\mathbf{Y}) + \sigma_\epsilon^2\mathbf{I}$$

* This is where we need GPs

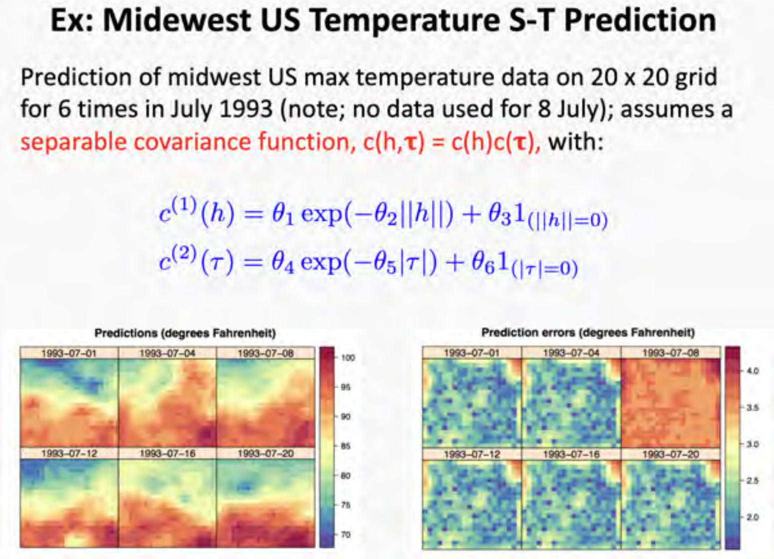
The universal kriging variance is given by:

Model-based uncertainty

$$\sigma^2_{Y,uk}(\mathbf{s}_0;t_0) = \sigma^2_Y - \mathbf{c}_0'\mathbf{C}_Z^{-1}\mathbf{c}_0 + \kappa$$

where $\kappa$ represents additional uncertainty brought to the prediction (relative to simple kriging) due to the estimation of the fixed effects.

# Brief Overview: Descriptive Spatio-Temporal Modeling



## Ex: Midewest US Temperature S-T Prediction

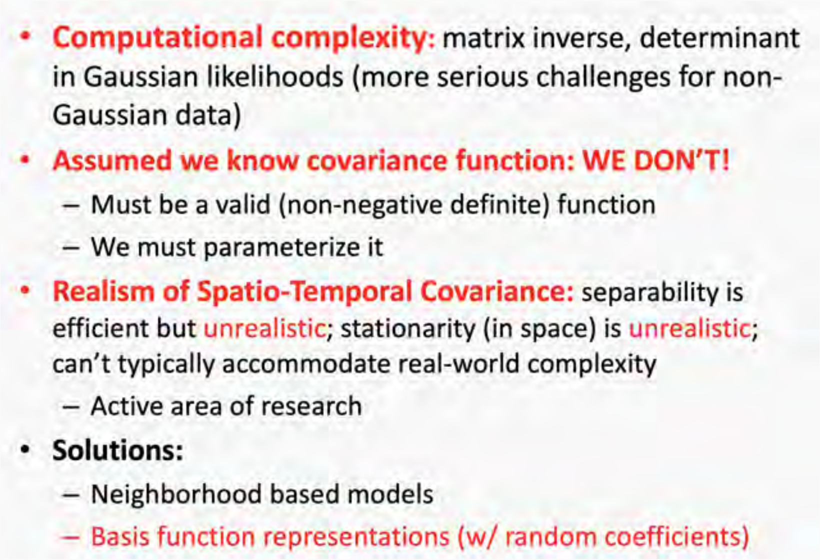
Prediction of midwest US max temperature data on 20 x 20 grid for 6 times in July 1993 (note; no data used for 8 July); assumes a separable covariance function, c(h,$\boldsymbol{\tau}$) = c(h)c($\boldsymbol{\tau}$), with:

$$c^{(1)}(h) = \theta_1 \exp(-\theta_2||h||) + \theta_3 1_{(||h||=0)}$$
$$c^{(2)}(\tau) = \theta_4 \exp(-\theta_5|\tau|) + \theta_6 1_{(|\tau|=0)}$$

Predictions (degrees Fahrenheit)

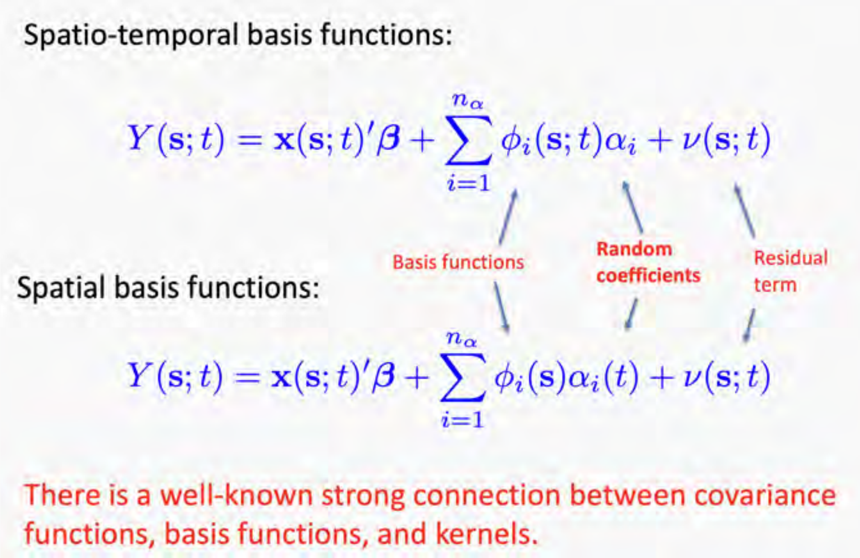
Prediction errors (degrees Fahrenheit)

# Brief Overview: Descriptive Spatio-Temporal Modeling



- **Computational complexity:** matrix inverse, determinant in Gaussian likelihoods (more serious challenges for non-Gaussian data)
- **Assumed we know covariance function: WE DON'T!**
  - Must be a valid (non-negative definite) function
  - We must parameterize it
- **Realism of Spatio-Temporal Covariance:** separability is efficient but unrealistic; stationarity (in space) is unrealistic; can't typically accommodate real-world complexity
  - Active area of research
- **Solutions:**
  - Neighborhood based models
  - Basis function representations (w/ random coefficients)

# Brief Overview: Descriptive Spatio-Temporal Modeling



Spatio-temporal basis functions:

$$Y(\mathbf{s};t) = \mathbf{x}(\mathbf{s};t)'\boldsymbol{\beta} + \sum_{i=1}^{n_\alpha} \phi_i(\mathbf{s};t)\alpha_i + \nu(\mathbf{s};t)$$

Basis functions

Random coefficients

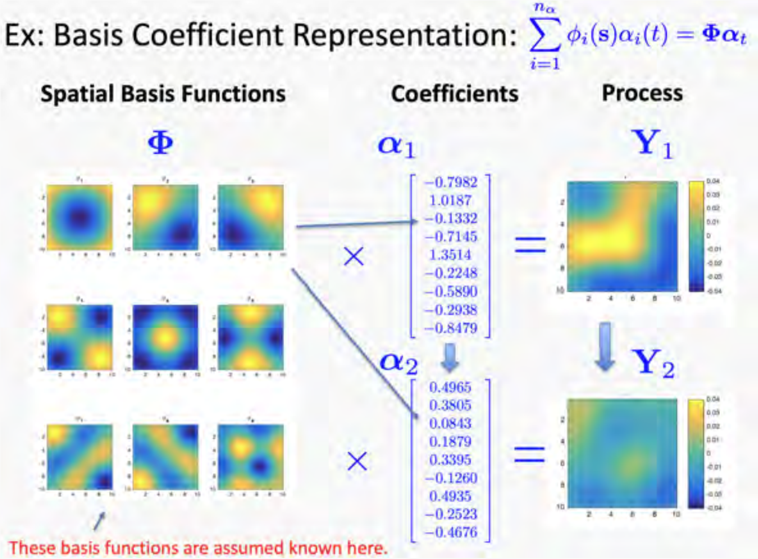
Residual term

Spatial basis functions:

$$Y(\mathbf{s};t) = \mathbf{x}(\mathbf{s};t)'\boldsymbol{\beta} + \sum_{i=1}^{n_\alpha} \phi_i(\mathbf{s})\alpha_i(t) + \nu(\mathbf{s};t)$$

There is a well-known strong connection between covariance functions, basis functions, and kernels.

# Brief Overview: Descriptive Spatio-Temporal Modeling



Ex: Basis Coefficient Representation: $\sum_{i=1}^{n_\alpha} \phi_i(\mathbf{s})\alpha_i(t) = \boldsymbol{\Phi}\boldsymbol{\alpha}_t$

**Spatial Basis Functions** **Coefficients** **Process**

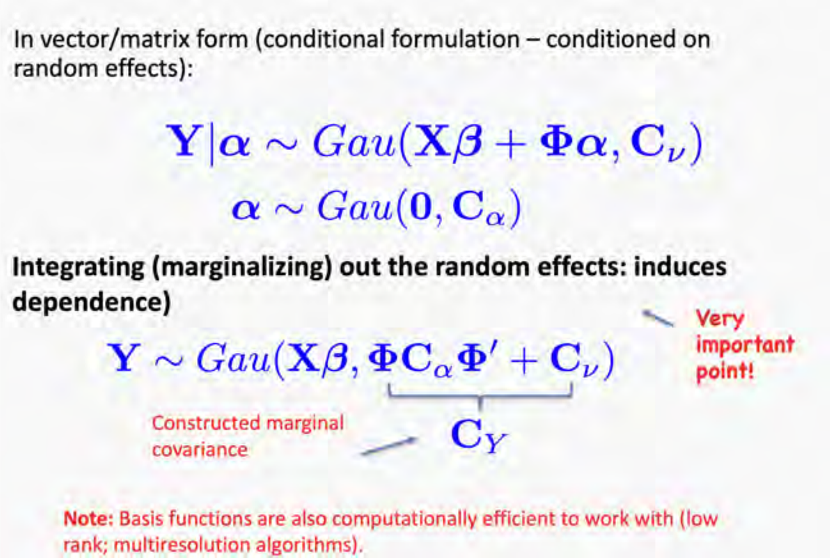
$\boldsymbol{\Phi}$

$$\times \quad \boldsymbol{\alpha}_1 \begin{bmatrix} -0.7982 \\ 1.0187 \\ -0.1332 \\ -0.7145 \\ 1.3514 \\ -0.2248 \\ -0.5890 \\ -0.2938 \\ -0.8479 \end{bmatrix} = \mathbf{Y}_1$$

$$\times \quad \boldsymbol{\alpha}_2 \begin{bmatrix} 0.4965 \\ 0.3805 \\ 0.0843 \\ 0.1879 \\ 0.3395 \\ -0.1260 \\ 0.4935 \\ -0.2523 \\ -0.4676 \end{bmatrix} = \mathbf{Y}_2$$

These basis functions are assumed known here.

# Brief Overview: Descriptive Spatio-Temporal Modeling



In vector/matrix form (conditional formulation – conditioned on random effects):

$$\mathbf{Y}|\boldsymbol{\alpha} \sim Gau(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Phi}\boldsymbol{\alpha}, \mathbf{C}_\nu)$$

$$\boldsymbol{\alpha} \sim Gau(\mathbf{0}, \mathbf{C}_\alpha)$$

**Integrating (marginalizing) out the random effects: induces dependence)**

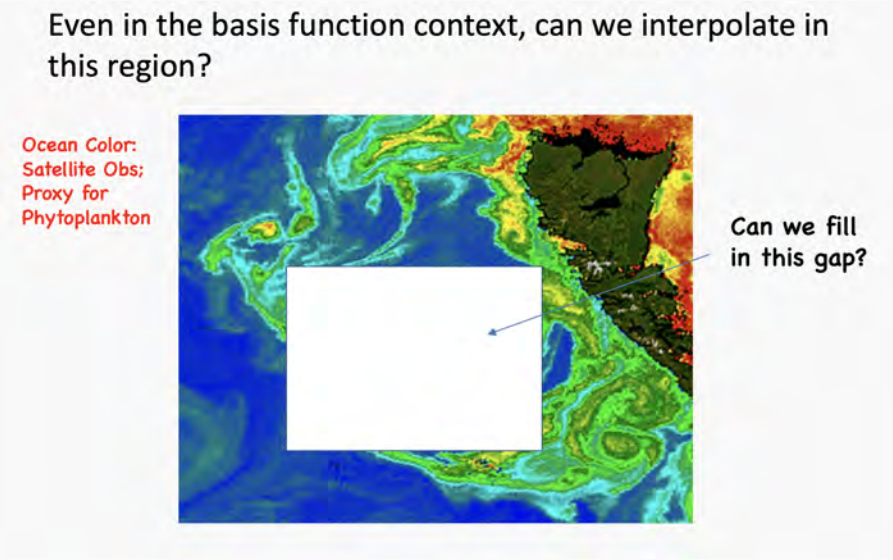
Very important point!

$$\mathbf{Y} \sim Gau(\mathbf{X}\boldsymbol{\beta}, \underbrace{\boldsymbol{\Phi}\mathbf{C}_\alpha\boldsymbol{\Phi}' + \mathbf{C}_\nu}_{\mathbf{C}_Y})$$

Constructed marginal covariance

**Note:** Basis functions are also computationally efficient to work with (low rank; multiresolution algorithms).

# Brief Overview: Descriptive Spatio-Temporal Modeling

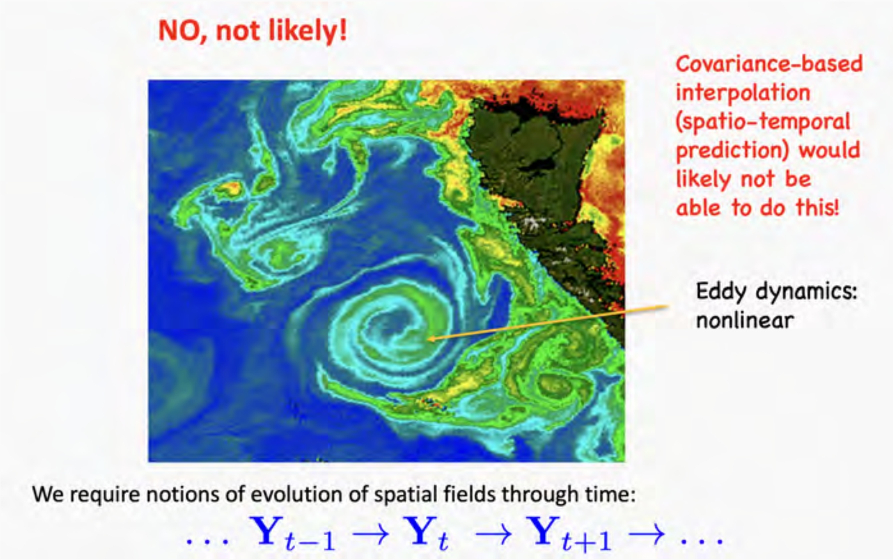


Even in the basis function context, can we interpolate in this region?

Ocean Color: Satellite Obs; Proxy for Phytoplankton

Can we fill in this gap?

# Brief Overview: Descriptive Spatio-Temporal Modeling



NO, not likely!

Covariance-based interpolation (spatio-temporal prediction) would likely not be able to do this!

Eddy dynamics: nonlinear

We require notions of evolution of spatial fields through time:

$$\ldots\ \mathbf{Y}_{t-1} \rightarrow \mathbf{Y}_{t} \rightarrow \mathbf{Y}_{t+1} \rightarrow \ldots$$

# Dynamic Models: Motivation

One of the biggest challenges with the *descriptive* approach to spatio-temporal modeling that we talked about before is that most real-world spatio-temporal processes are more complex than can be realistically specified by the relatively simple classes of spatio-temporal covariance functions that are available.

We can improve on this by using basis-function expansions and random effects, but it is often the case that the random effects have temporal dependence.

We can accommodate this temporal dependence by considering a *conditional* or *dynamic* perspective, in which we think of spatial processes (or basis coefficients) evolving through time in some probabilistic manner.

# Dynamics

- Spatio-temporal dynamics are due to the interaction of the process components across space and time and/or across scales of variability
    - Some types of interaction make sense for some processes, and some don't (e.g., process knowledge should not be ignored if available)
    - Statisticians have often ignored such knowledge!
- Dimensionality can prevent the (efficient) estimation of model parameters
    - Requires sensible science-based parameterizations and/or dimension reduction; sparse structures; regularization
    - Deep (hierarchical) representations can help here as well

# Spatio-Temporal Modeling

**Dynamic (Conditional) Approach:**

Current values of the process at a location evolve from past values of the process at various locations

- Need only specify conditional distributions
- Closer to the etiology of the phenomenon under study
- More likely to establish answers to the "why" question (causality) – better for forecasting and prediction in big gaps
- Are best when there is some *a priori* knowledge available concerning process behavior (i.e., mechanistic behavior)
- Study of how things change over time
- Pattern of change or growth of a system over time

# September 25, 2010: Radar Nowcasting Application!

It may look like something out of a movie, but it's not — this was the scene from early Saturday afternoon in Kansas. (AP Photo/Charlie Riedel)

# Basic Modeling Framework

There are two critical assumptions for DSTMs:

- Data conditioned on the latent process can be factored into the product of independent distributions:

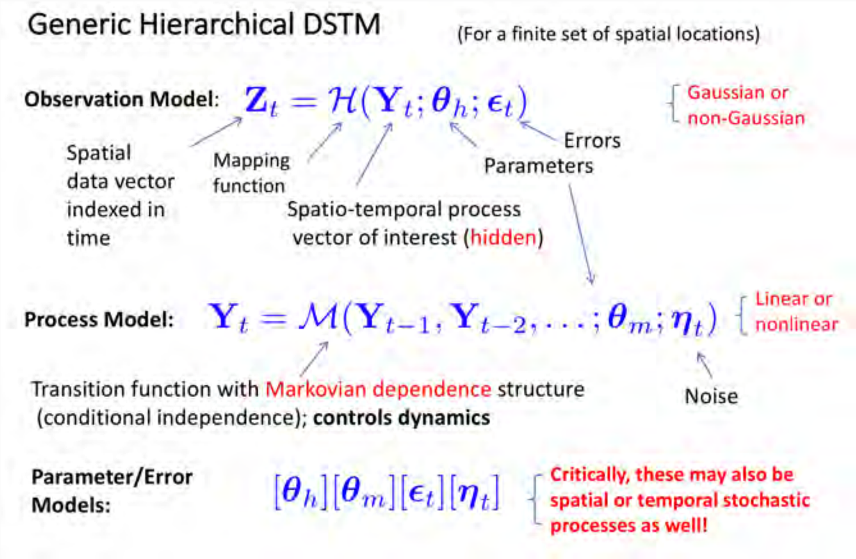
$$[\mathbf{z}_T, \ldots, \mathbf{z}_1 | \mathbf{Y}_T, \ldots, \mathbf{Y}_1, \boldsymbol{\theta}_d] = \prod_{t=1}^{T} [\mathbf{z}_t | \mathbf{Y}_t, \boldsymbol{\theta}_d]$$

- The joint distribution of the latent process can be factored into conditional (in time) models (e.g., first-order model):

$$[\mathbf{Y}_T, \ldots, \mathbf{Y}_1, \mathbf{Y}_0 | \boldsymbol{\theta}_p] = \prod_{t=1}^{T} [\mathbf{Y}_t | \mathbf{Y}_{t-1}, \boldsymbol{\theta}_p][\mathbf{Y}_0 | \boldsymbol{\theta}_p]$$

Challenge: specification of the models associated with these component distributions

# General Hierarchical DSTM



## Generic Hierarchical DSTM

(For a finite set of spatial locations)

**Observation Model:** $\mathbf{Z}_t = \mathcal{H}(\mathbf{Y}_t; \boldsymbol{\theta}_h; \epsilon_t)$

Gaussian or non-Gaussian

Spatial data vector indexed in time

Mapping function

Spatio-temporal process vector of interest (hidden)

Parameters

Errors

**Process Model:** $\mathbf{Y}_t = \mathcal{M}(\mathbf{Y}_{t-1}, \mathbf{Y}_{t-2}, \ldots; \boldsymbol{\theta}_m; \boldsymbol{\eta}_t)$

Linear or nonlinear

Transition function with Markovian dependence structure (conditional independence); **controls dynamics**

Noise

**Parameter/Error Models:** $[\boldsymbol{\theta}_h][\boldsymbol{\theta}_m][\epsilon_t][\boldsymbol{\eta}_t]$

**Critically, these may also be spatial or temporal stochastic processes as well!**

# Linear DSTMs: Process Model

We need a model corresponding to the distribution:

$$[\mathbf{Y}_t | \mathbf{Y}_{t-1}, \mathbf{Y}_{t-2}, \ldots, \boldsymbol{\theta}_m] \quad \text{(often assume } [\mathbf{Y}_t | \mathbf{Y}_{t-1}, \boldsymbol{\theta}_m]\text{)}$$

Assumption: the process at the current time is related to the process at a previous time (or times). In general, we can consider spatio-temporal processes in continuous time and/or space. For brevity, we focus here on the case where time is discrete and equally spaced.

We refer to such a process as a space-time dynamical process.

# Linear DSTM: Process Model

In the linear process case, the conditional distribution $[\mathbf{Y}_t | \mathbf{Y}_{t-1}, \boldsymbol{\theta}_m]$ implies a first-order Markov model of the form:

$$Y_t(\mathbf{s}_i) = \sum_{j=1}^{n} m_{ij}(\boldsymbol{\theta}_m) Y_{t-1}(\mathbf{s}_j) + \eta_t(\mathbf{s}_i),$$

where $m_{ij}(\boldsymbol{\theta}_m)$ describes how the spatial process at the previous time at location $j$ gets redistributed to location $i$ at the current time, according to the parameters, $\boldsymbol{\theta}$

These equations imply a matrix model:

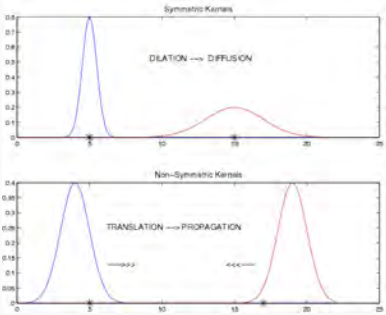
$$\mathbf{Y}_t = \mathbf{M}(\boldsymbol{\theta}_m)\mathbf{Y}_{t-1} + \boldsymbol{\eta}_t, \;\; \boldsymbol{\eta}_t \sim Gau(\mathbf{0}, \mathbf{C}_\eta)$$

**Challenge:** parsimonious parameterization of $\mathbf{M}(\boldsymbol{\theta}_m)$!

# Linear DSTM: Process Behavior

Dynamical behavior is implied by changes in the transition operator "shape": e.g., linear spatio-temporal processes often exhibit advective and diffusive behavior:

- "width" (decay rate) of the transition operator neighborhood controls the rate of spread (diffusion)
- degree of "asymmetry" in the transition operator controls the speed and direction of propagation (advection)
- "long range dependence" can be accommodated by "multimodal" operators and/or heavy tails

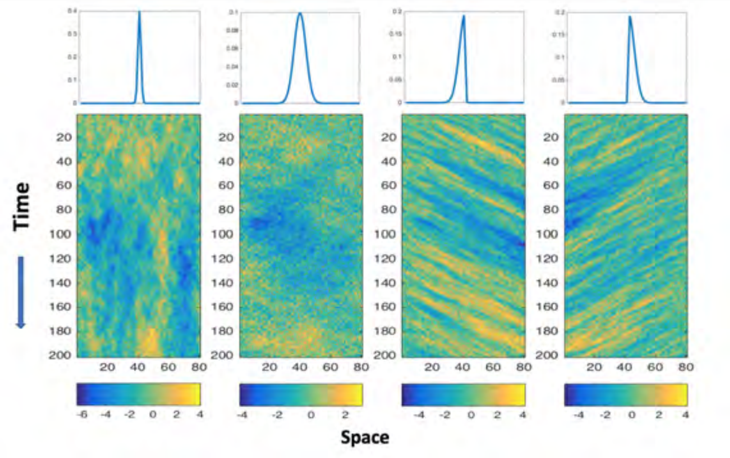


This suggests ways that we might parameterize the transition operator and/or induce sparse structure.

# Linear DSTM: Process Behavior

Integro-difference equation (IDE) (kernel) representation:

$$Y_t(\mathbf{s}) = \int_{D_s} m(\mathbf{s}, \mathbf{x}; \boldsymbol{\theta}_m) Y_{t-1}(\mathbf{x}) d\mathbf{x} + \eta_t(\mathbf{s}), \quad \mathbf{s}, \mathbf{x} \in D_s$$



# Basic Hierarchical Linear DSTM

**Data:**

$$\mathbf{z}_t = \mathbf{H}_t \mathbf{Y}_t + \boldsymbol{\epsilon}_t, \;\; \boldsymbol{\epsilon}_t \sim Gau(\mathbf{0}, \mathbf{C}_{\epsilon,t}(\boldsymbol{\theta}_d))$$

**Process:**

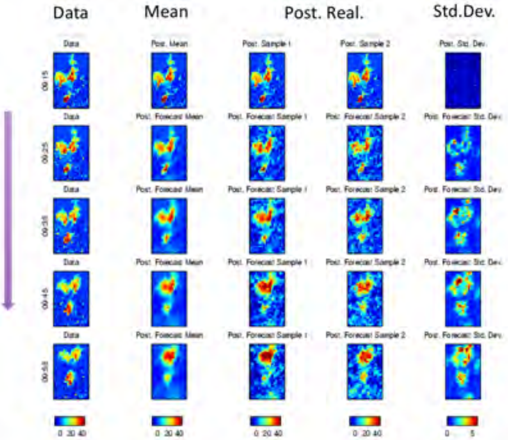
$$\mathbf{Y}_t = \mathbf{M}(\boldsymbol{\theta}_{m,1}) \mathbf{Y}_{t-1} + \boldsymbol{\eta}_t, \;\; \boldsymbol{\eta}_t \sim Gau(\mathbf{0}, \mathbf{C}_{\eta}(\boldsymbol{\theta}_{m,2}))$$

**Parameters:**

$\boldsymbol{\theta}_d, \; \boldsymbol{\theta}_{m,1}, \; \boldsymbol{\theta}_{m,2}$

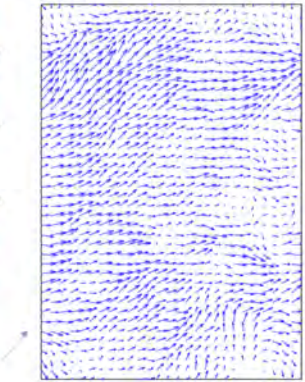
These parameters may be estimated empirically, but we get more flexibility if they are given dependent prior distributions, such as Gaussian random process priors (that may depend on other variables), and they can easily be allowed to vary with time and/or space.

# Example: Radar Nowcasting (Sydney, pre 2000 Olympics)



Statistical model motivated by a linear advection-diffusion process with spatially varying parameters.

Implied Propagation by **M**



Xu et al. (2005; JASA)

Depicts skewness and relative decay of transition weights

# Nonlinear Spatio-Temporal Processes

Many real-world processes are not linear (e.g., growth, nonlinear advection, density dependence, shock waves, repulsion, predator-prey, etc.)

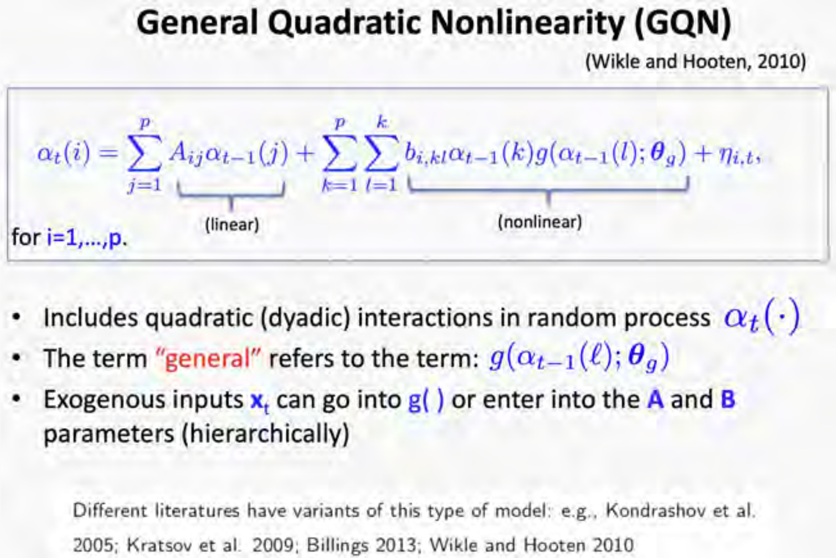
- Nonlinear dynamical behavior arises from the complicated interactions across spatio-temporal scales of variability and interactions across multiple processes
- It is important to consider this if forecasting or filling in big gaps is important
- Examples in mechanistic models across many disciplines

# Nonlinear DSTMs: Statistical Approaches

Over the past few years, several parametric and non-parametric statistical approaches have proven useful for hierarchical formulations of nonlinear DSTMs. Some of these have been motivated by non-statistical models from other disciplines. It is still early in the development of these methods.

- Time-Varying Parameters
- General Quadratic Nonlinearity (GQN)
- Individual (Agent)-Based Models
- “Mechanism-Free” Analog Embedding Models
- Neural Network Models

# Nonlinear Spatio-Temporal Dynamic Models



## General Quadratic Nonlinearity (GQN)

(Wikle and Hooten, 2010)

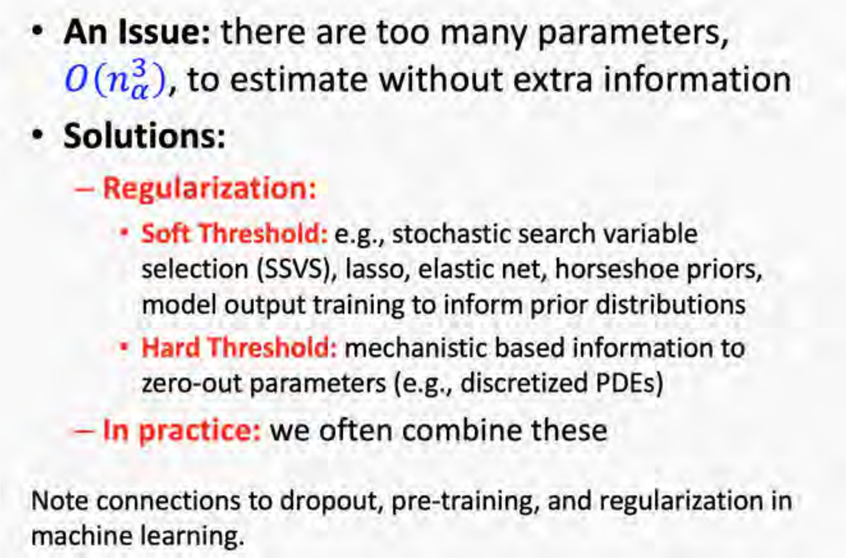
$$\alpha_t(i) = \sum_{j=1}^{p} \underbrace{A_{ij}\alpha_{t-1}(j)}_{\text{(linear)}} + \sum_{k=1}^{p}\sum_{l=1}^{k} \underbrace{b_{i,kl}\alpha_{t-1}(k)g(\alpha_{t-1}(l);\boldsymbol{\theta}_g)}_{\text{(nonlinear)}} + \eta_{i,t},$$

for i=1,...,p.

- Includes quadratic (dyadic) interactions in random process $\alpha_t(\cdot)$
- The term "general" refers to the term: $g(\alpha_{t-1}(\ell);\boldsymbol{\theta}_g)$
- Exogenous inputs $\mathbf{x}_t$ can go into g( ) or enter into the **A** and **B** parameters (hierarchically)

Different literatures have variants of this type of model: e.g., Kondrashov et al. 2005; Kratsov et al. 2009; Billings 2013; Wikle and Hooten 2010

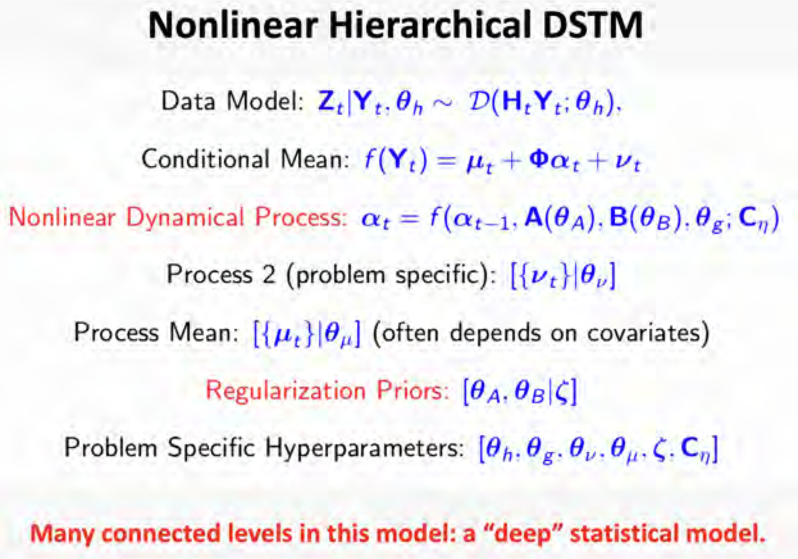
# Nonlinear Spatio-Temporal Dynamic Models



- **An Issue:** there are too many parameters, $O(n_\alpha^3)$, to estimate without extra information
- **Solutions:**
  - **Regularization:**
    - **Soft Threshold:** e.g., stochastic search variable selection (SSVS), lasso, elastic net, horseshoe priors, model output training to inform prior distributions
    - **Hard Threshold:** mechanistic based information to zero-out parameters (e.g., discretized PDEs)
  - **In practice:** we often combine these

Note connections to dropout, pre-training, and regularization in machine learning.

# Nonlinear Spatio-Temporal Dynamic Models



## Nonlinear Hierarchical DSTM

Data Model: $\mathbf{Z}_t|\mathbf{Y}_t, \boldsymbol{\theta}_h \sim \mathcal{D}(\mathbf{H}_t\mathbf{Y}_t; \boldsymbol{\theta}_h)$.

Conditional Mean: $f(\mathbf{Y}_t) = \boldsymbol{\mu}_t + \boldsymbol{\Phi}\boldsymbol{\alpha}_t + \boldsymbol{\nu}_t$

Nonlinear Dynamical Process: $\boldsymbol{\alpha}_t = f(\boldsymbol{\alpha}_{t-1}, \mathbf{A}(\boldsymbol{\theta}_A), \mathbf{B}(\boldsymbol{\theta}_B), \boldsymbol{\theta}_g; \mathbf{C}_\eta)$

Process 2 (problem specific): $[\{\boldsymbol{\nu}_t\}|\boldsymbol{\theta}_\nu]$

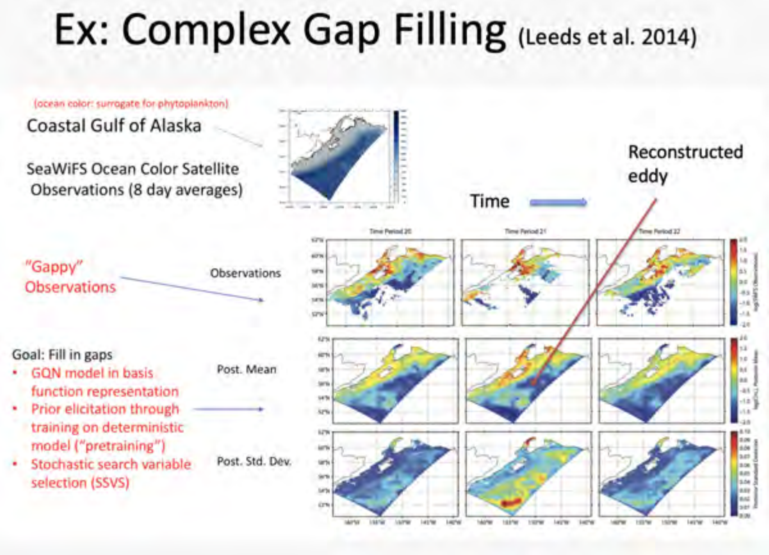
Process Mean: $[\{\boldsymbol{\mu}_t\}|\boldsymbol{\theta}_\mu]$ (often depends on covariates)

Regularization Priors: $[\boldsymbol{\theta}_A, \boldsymbol{\theta}_B|\zeta]$

Problem Specific Hyperparameters: $[\boldsymbol{\theta}_h, \boldsymbol{\theta}_g, \boldsymbol{\theta}_\nu, \boldsymbol{\theta}_\mu, \zeta, \mathbf{C}_\eta]$

**Many connected levels in this model: a "deep" statistical model.**

# Nonlinear Spatio-Temporal Dynamic Models



## Ex: Complex Gap Filling (Leeds et al. 2014)

(ocean color: surrogate for phytoplankton)

Coastal Gulf of Alaska

SeaWiFS Ocean Color Satellite Observations (8 day averages)

Time

Reconstructed eddy

"Gappy" Observations

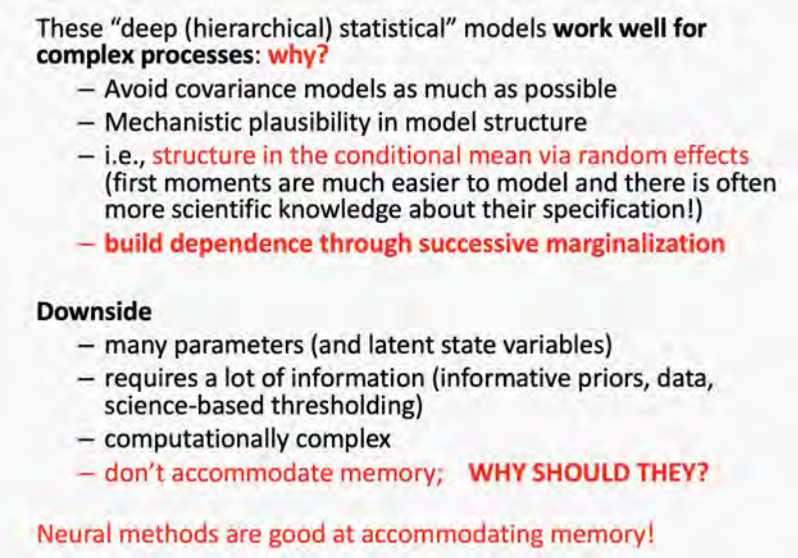
Observations

Goal: Fill in gaps
- GQN model in basis function representation
- Prior elicitation through training on deterministic model ("pretraining")
- Stochastic search variable selection (SSVS)

Post. Mean

Post. Std. Dev.

# Nonlinear Spatio-Temporal Dynamic Models



These “deep (hierarchical) statistical” models **work well for complex processes**: **why?**

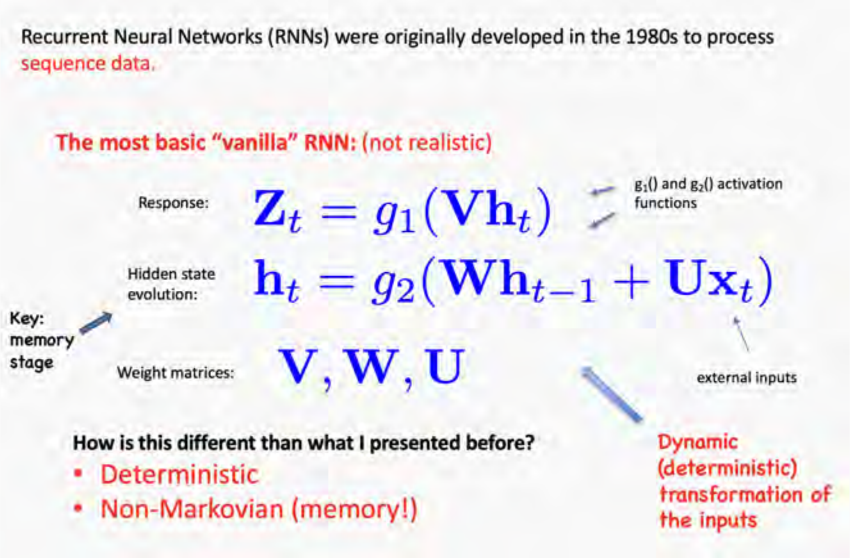
- Avoid covariance models as much as possible
- Mechanistic plausibility in model structure
- i.e., structure in the conditional mean via random effects (first moments are much easier to model and there is often more scientific knowledge about their specification!)
- **build dependence through successive marginalization**

**Downside**

- many parameters (and latent state variables)
- requires a lot of information (informative priors, data, science-based thresholding)
- computationally complex
- don’t accommodate memory; **WHY SHOULD THEY?**

Neural methods are good at accommodating memory!

# Nonlinear Spatio-Temporal Dynamic Models: RNN



Recurrent Neural Networks (RNNs) were originally developed in the 1980s to process sequence data.

**The most basic "vanilla" RNN:** (not realistic)

Response: $$\mathbf{Z}_t = g_1(\mathbf{V}\mathbf{h}_t)$$

$g_1$() and $g_2$() activation functions

Hidden state evolution: $$\mathbf{h}_t = g_2(\mathbf{W}\mathbf{h}_{t-1} + \mathbf{U}\mathbf{x}_t)$$

Key: memory stage

Weight matrices: $$\mathbf{V}, \mathbf{W}, \mathbf{U}$$

external inputs

**How is this different than what I presented before?**

- Deterministic
- Non-Markovian (memory!)

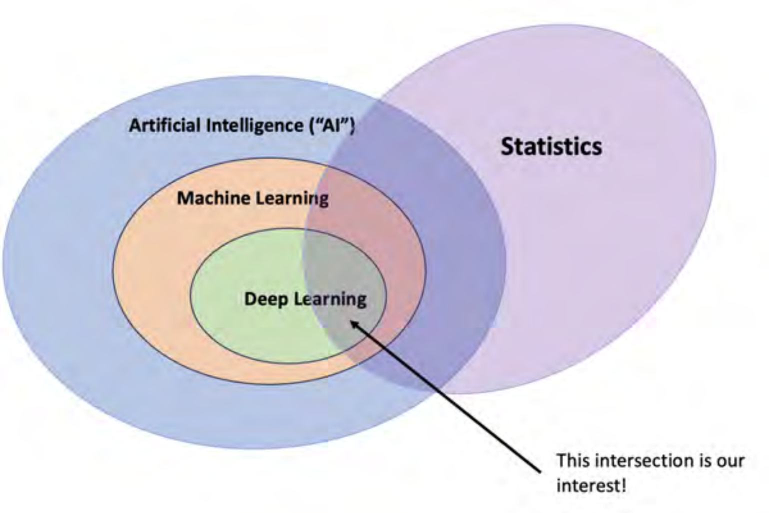
Dynamic (deterministic) transformation of the inputs

# Nonlinear Spatio-Temporal Dynamic Models: RNN

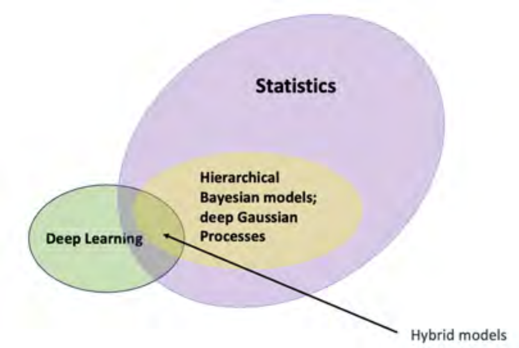
- This traditional RNN is very difficult to train (vanishing/exploding gradient problem)
- Alternatives:
  - Long Short-Term Memory (LSTM) Models
  - Gated Recurrent Unit (GRU) Models
  - Echo State Networks (reservoir computing; random weights)
- Issues: uncertainty quantification and interpretability

Consider hybrid statistical/neural models

# Hybrid Neural/Statistical Models

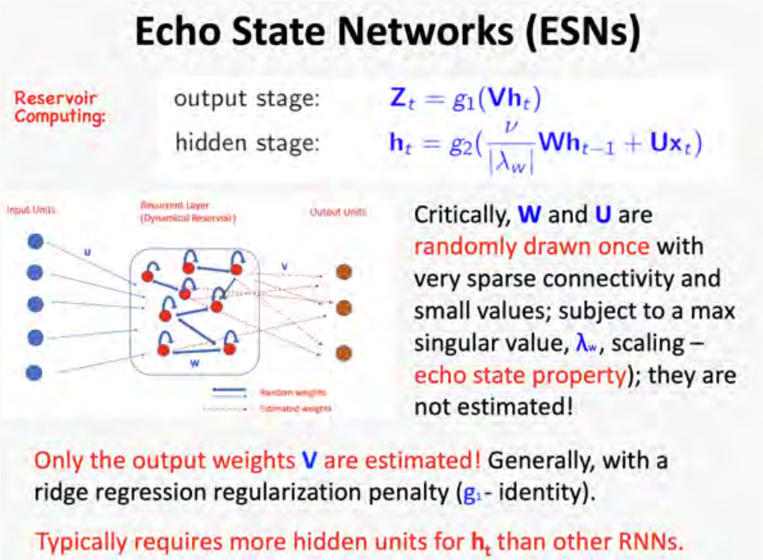

Artificial Intelligence ("AI")
Machine Learning
Deep Learning
Statistics
This intersection is our interest!

# Hybrid Neural/Statistical Models



- Hybrid models have recently been developed for latent process, data, and parameter specifications.
  - integration of statistical modeling ideas with deep neural network models to take advantage of the strengths of each for complex spatio-temporal data (see Wikle and Zammit-Mangion, 2023)

# Nonlinear Spatio-Temporal Dynamic Models: ESN



## Echo State Networks (ESNs)

Reservoir Computing:

output stage: $\mathbf{Z}_t = g_1(\mathbf{V}\mathbf{h}_t)$

hidden stage: $\mathbf{h}_t = g_2(\frac{\nu}{|\lambda_w|}\mathbf{W}\mathbf{h}_{t-1} + \mathbf{U}\mathbf{x}_t)$

Critically, **W** and **U** are randomly drawn once with very sparse connectivity and small values; subject to a max singular value, $\lambda_w$, scaling – echo state property); they are not estimated!

Only the output weights **V** are estimated! Generally, with a ridge regression regularization penalty ($g_1$ - identity).

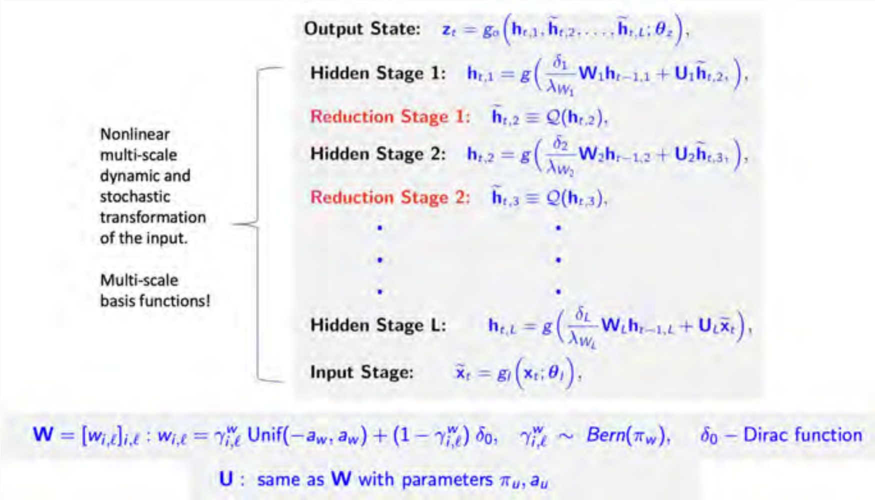
Typically requires more hidden units for $\mathbf{h}_t$ than other RNNs.

# Nonlinear Spatio-Temporal Dynamic Models: ESN

ESN models have been used in several spatio-temporal statistical analyses, combined with ensemble-based uncertainty quantification:

- Long-lead forecasting of El Niño (McDermott and Wikle, 2017)
- Long-lead forecasting of soil moisture using deep ESNs (McDermott and Wikle, 2019)
- Forecasting the spread of wildfires (Yoo and Wikle, 2023)
- Forecasting urban air pollution (Bonas and Castruccio, 2023)
- Forecasting epidemiological spread; raccoon rabies spread (Grieshop and Wikle, 2023)

# Nonlinear Spatio-Temporal Dynamic Models: Deep ESN



Nonlinear multi-scale dynamic and stochastic transformation of the input.

Multi-scale basis functions!

**Output State:** $\mathbf{z}_t = g_o\left(\mathbf{h}_{t,1}, \tilde{\mathbf{h}}_{t,2}, \ldots, \tilde{\mathbf{h}}_{t,L}; \boldsymbol{\theta}_z\right),$

**Hidden Stage 1:** $\mathbf{h}_{t,1} = g\left(\frac{\delta_1}{\lambda_{W_1}} \mathbf{W}_1 \mathbf{h}_{t-1,1} + \mathbf{U}_1 \tilde{\mathbf{h}}_{t,2},\right),$

**Reduction Stage 1:** $\tilde{\mathbf{h}}_{t,2} \equiv \mathcal{Q}(\mathbf{h}_{t,2}),$

**Hidden Stage 2:** $\mathbf{h}_{t,2} = g\left(\frac{\delta_2}{\lambda_{W_2}} \mathbf{W}_2 \mathbf{h}_{t-1,2} + \mathbf{U}_2 \tilde{\mathbf{h}}_{t,3},\right),$

**Reduction Stage 2:** $\tilde{\mathbf{h}}_{t,3} \equiv \mathcal{Q}(\mathbf{h}_{t,3}),$

$\vdots \qquad \vdots$

**Hidden Stage L:** $\mathbf{h}_{t,L} = g\left(\frac{\delta_L}{\lambda_{W_L}} \mathbf{W}_L \mathbf{h}_{t-1,L} + \mathbf{U}_L \tilde{\mathbf{x}}_t\right),$

**Input Stage:** $\tilde{\mathbf{x}}_t = g_I\left(\mathbf{x}_t; \boldsymbol{\theta}_I\right),$

$\mathbf{W} = [w_{i,\ell}]_{i,\ell} : w_{i,\ell} = \gamma^{w}_{i,\ell}\ \text{Unif}(-a_w, a_w) + (1 - \gamma^{w}_{i,\ell})\ \delta_0, \quad \gamma^{w}_{i,\ell} \sim \textit{Bern}(\pi_w), \quad \delta_0$ – Dirac function

$\mathbf{U}$ : same as $\mathbf{W}$ with parameters $\pi_u, a_u$

McDermott and Wikle (2019)

# Nonlinear Spatio-Temporal Dynamic Models: ESN

## 6 mo. Forecast of Soil Moisture over US Corn Belt

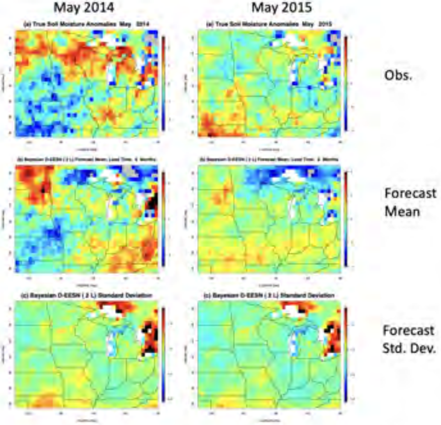
**3-Level Deep Ensemble ESN Model:**

6 mo forecast of monthly soil moisture anomalies (15 Spatial-Temporal PCs) given lagged SST anomalies (5 PCs) as input (different time scales)

Training: 1948 – 2011

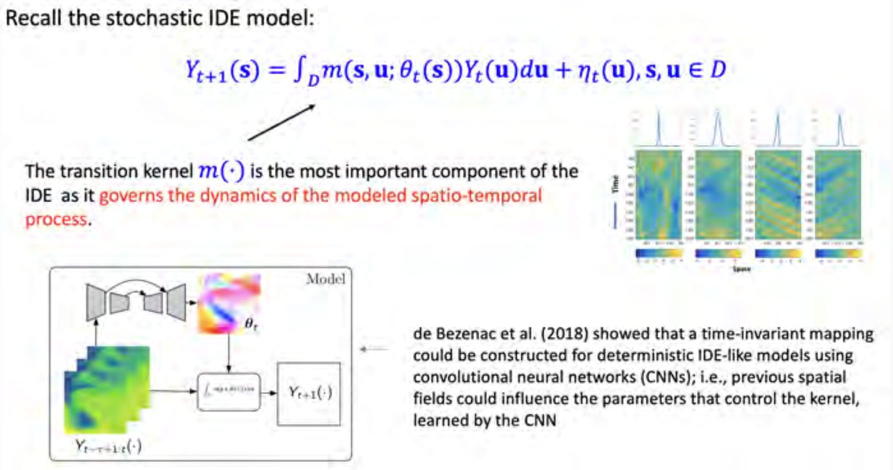
Out-of-sample forecasts from November:
May 2012 – 2017

Focus on May forecasts given it is a crucial time in corn phenology


May 2014
May 2015
(a) True Soil Moisture Anomalies May 2014
(a) True Soil Moisture Anomalies May 2015
Obs.
(b) Bayesian D-EESN (3 L) Forecast Mean; Lead Time: 6 Months
(b) Bayesian D-EESN (3 L) Forecast Mean; Lead Time: 6 Months
Forecast Mean
(c) Bayesian D-EESN (3 L) Standard Deviation
(c) Bayesian D-EESN (3 L) Standard Deviation
Forecast Std. Dev.

McDermott and Wikle, 2019

# Nonlinear Spatio-Temporal Dynamic Models: Hybrid IDE/CNN Model



Recall the stochastic IDE model:

$$Y_{t+1}(\mathbf{s}) = \int_D m(\mathbf{s},\mathbf{u};\theta_t(\mathbf{s}))Y_t(\mathbf{u})d\mathbf{u} + \eta_t(\mathbf{u}), \mathbf{s},\mathbf{u} \in D$$

The transition kernel $m(\cdot)$ is the most important component of the IDE as it governs the dynamics of the modeled spatio-temporal process.

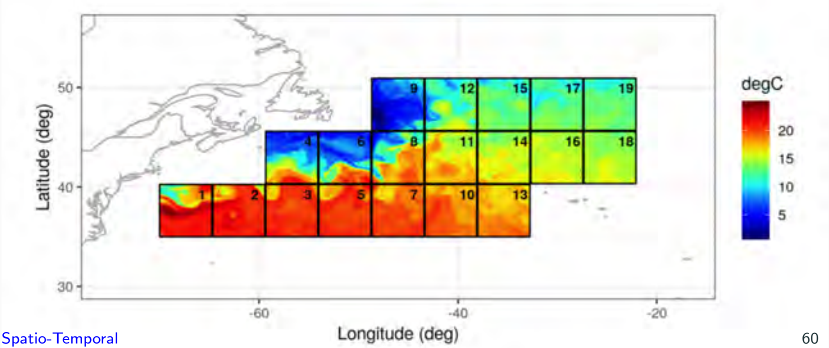
de Bezenac et al. (2018) showed that a time-invariant mapping could be constructed for deterministic IDE-like models using convolutional neural networks (CNNs); i.e., previous spatial fields could influence the parameters that control the kernel, learned by the CNN

# Nonlinear Spatio-Temporal Dynamic Models: Hybrid IDE/CNN Model

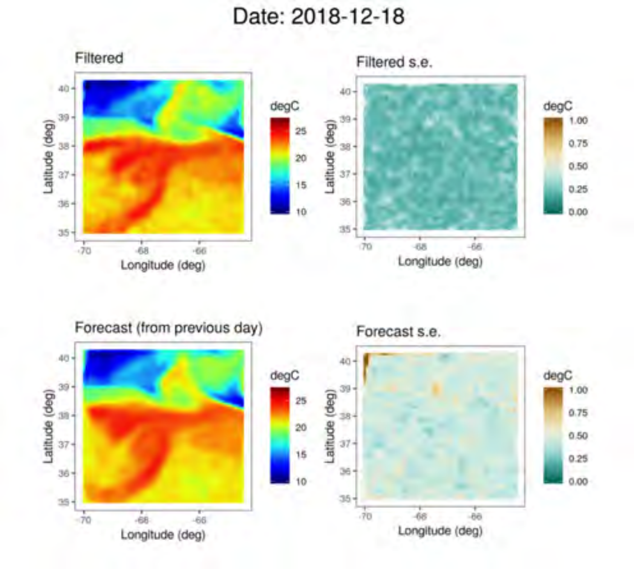
- CNNs can work well in this setting because they encode "dynamical features" that can have a straight-forward relationship to IDE parameters (i.e., advection and diffusion)
- However, CNNs are designed to work on noiseless, complete, images.
- Zammit-Mangion and Wikle (2020) extend the de Bezenac et al. approach to the case where data are noisy and incomplete and where the IDE model and parameters are uncertain.
- This is what hierarchical DSTMs are good at; i.e., suggesting a hybrid approach

# IDE/CNN: SST Example

- Daily SST (1/12 degree lat/lon grid) over 19 zones (each 64 × 64) in the North Atlantic
- First 4003 days (12/27/2006 to 12/11/2017) used for CNN training; $\tau = 3$; $\approx 2,000,000$ CNN parameters trained by stochastic gradient descent on GPUs
- Evaluate on daily forecasts from 9/1/2018 - 12/31/2018



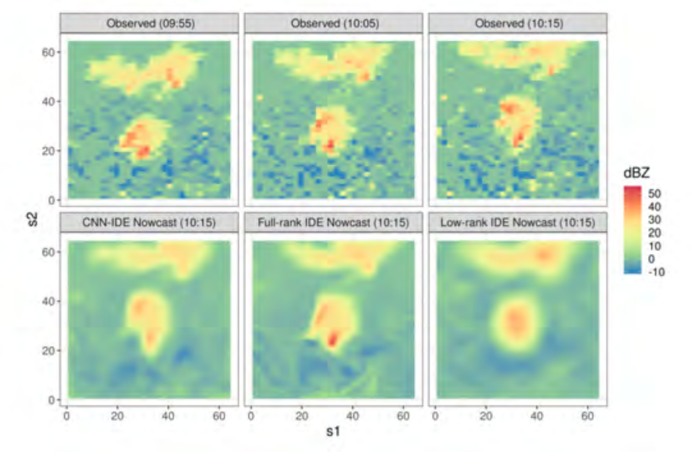
# IDE/CNN: SST Results



# IDE/CNN: “Transfer Learning”

- If the IDE/CNN encodes the physical information, can we use it for forecasting in entirely different applications?
- YES! – If the process is governed by similar underlying physical principles (e.g., advection, diffusion, etc.)
- We did an unusual experiment where we used the IDE/CNN trained on SST data to forecast (nowcast) radar reflectivity 10 minutes into the future from 11 sequential images of radar precipitation (e.g., see Xu et al. 2005)
- We compared once again to the time-varying IDE in a sliding window, and a low-rank IDE from the R package IDE.

# IDE/CNN: Performance on Completely Different Process



Radar-reflectivity images at 09:55, 10:05, and 10:15 UTC and nowcasts for 10:15 UTC.

# Hybrid Models

The examples presented here are hybrid neural/statistical models where the neural approach is embedded within a statistical model to improve prediction.

Another type of hybrid model concerns the case where an advanced statistical model is embedded within a machine learning/neural framework.

**Example:** embedding a flexible spatial extremes model within a variational autoencoder to emulate spatial processes with dependent extremes (e.g., Zhang et al. 2023; forthcoming).

# Future of Spatio-Temporal Statistics

- High-dimensions – dealing with redundancy (simultaneously we can have too much and too little data!)
- Data and processes with different spatio-temporal support (resolution)
- Multiple spatio-temporal processes interacting (nonlinearly)
- Hybrid machine learning/statistics methods with uncertainty quantification and interpretability
- Combinations of models that consider prediction, inference, classification
- Simplified implementations and programming

# References

Bonas, M., and Castruccio, S., 2023: Calibration of Spatio-Temporal Forecasts from Citizen Science Urban Air Pollution Data with Sparse Recurrent Neural Networks, Annals of Applied Statistics, in press.

de Bézenac, E., Pajot, A., & Gallinari, P., 2019: Deep learning for physical processes: Incorporating prior scientific knowledge. Journal of Statistical Mechanics: Theory and Experiment, 2019(12), 124009.

Grieshop, N., & Wikle, C. K., 2023: Bayesian Ensemble Echo State Networks for Enhancing Binary Stochastic Cellular Automata. arXiv preprint arXiv:2306.04696.

Leeds, W.B., Wikle, C.K., & J. Fiechter, 2014: Emulator-assisted reduced-rank ecological data assimilation for multivariate dynamical spatio-temporal processes. Statistical Methodology,17, 126-138.

McDermott, P.L. and C.K. Wikle, 2019: Deep echo state networks with uncertainty quantification for spatio-temporal forecasting. Environmetrics, 30, https://doi.org/10.1002/env.2553.

McDermott, P.L., and C.K. Wikle, 2017: An ensemble quadratic echo state network for nonlinear spatio- temporal forecasting. Stat, 6, 315–330, doi:10.1002/sta4.160.

Yoo, M., & Wikle, C. K., 2023: Using echo state networks to inform physical models for fire front propagation. *Spatial Statistics*, 54, 100732.

Wikle, C.K. & M.B. Hooten, 2010: A general science-based framework for spatio-temporal dynamical models. Invited discussion paper for Test, 19, 417-451.

# References

Wikle, C.K., Zammit-Mangion, A., & Cressie, N., 2019: Spatio-Temporal Statistics with R. Chapman & Hall/CRC.

Wikle, C.K. & A. Zammit-Mangion, 2023: Statistical deep learning for spatial and spatiotemporal data. Annual Review Statistics and its Application. 10:247-270

Xu, K., Wikle, C. K., & Fox, N. I., 2005: A kernel-based spatio-temporal dynamical model for nowcasting weather radar reflectivities. Journal of the American statistical Association, 100(472), 1133-1144.

Zammit-Mangion, A., & Wikle, C. K., 2020: Deep integro-difference equation models for spatio-temporal forecasting. Spatial Statistics, 37, 100408.

Zhang, L, Ma, X., Wikle, C.K., & R. Huser, 2023: Flexible and efficient spatial extremes emulation via variational autoencoders. arXiv:2307.08079.