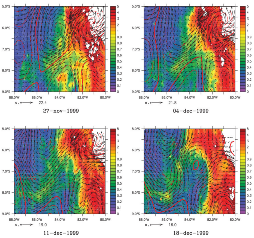
# Empirical Orthogonal Functions and derived methods

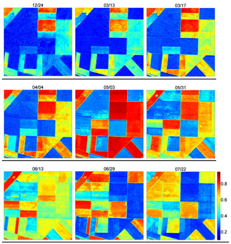
B. Alglave, B. Dufee, S. Obakrim and J. Thorson

March 2024

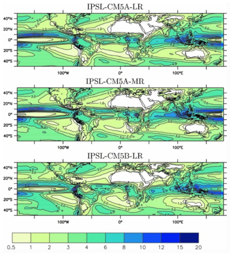
## Oceanography



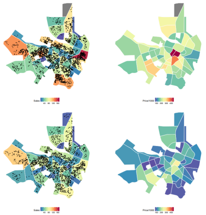

## Agronomy



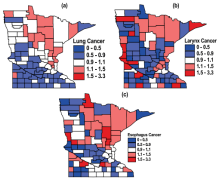

## Climatology



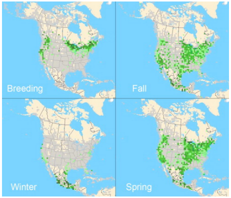

## Economics



## Epidemiology




## Ecology



# Empirical Orthogonal Functions

First introduced by Lorenz (1956)

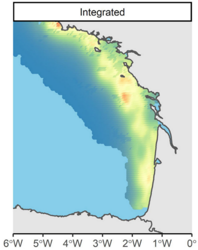
➠ What is the **best representation** for a spatio-temporal field ?

➠ How to perform **dimension-reduction** on a spatio-temporal field ?

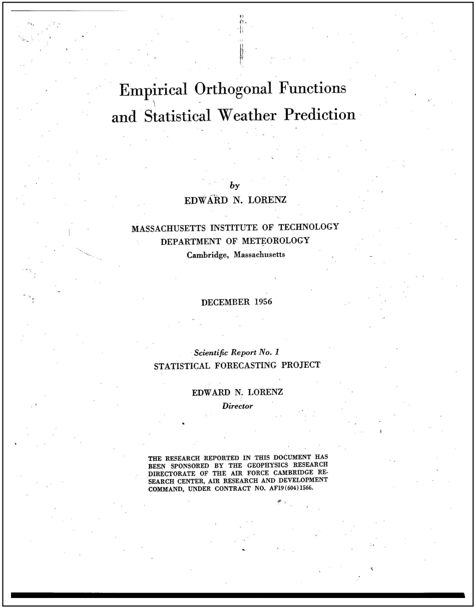
➠ How to make **projections**?

- - - - - - - - - - - - - - - - - - - - - - - - - - -

Presentation based on a fishery case study.
Demersal fisheries of the Bay of Biscay.
Monthly maps from 2008 to 2018 (132 maps)




**Solea solea**



Empirical Orthogonal Functions
and Statistical Weather Prediction

*by*
EDWARD N. LORENZ

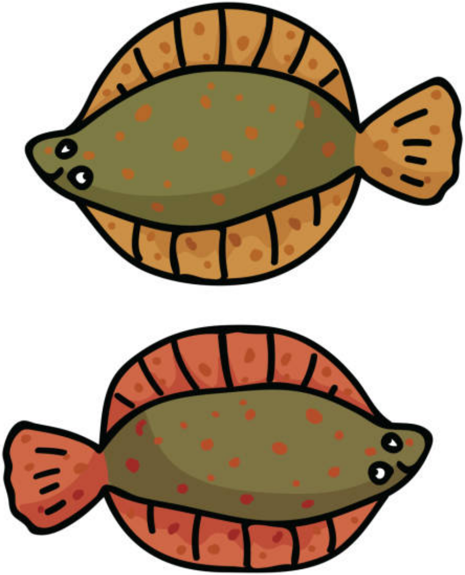
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
DEPARTMENT OF METEOROLOGY
Cambridge, Massachusetts

DECEMBER 1956

*Scientific Report No. 1*
STATISTICAL FORECASTING PROJECT

EDWARD N. LORENZ
*Director*

THE RESEARCH REPORTED IN THIS DOCUMENT HAS BEEN SPONSORED BY THE GEOPHYSICS RESEARCH DIRECTORATE OF THE AIR FORCE CAMBRIDGE RESEARCH CENTER, AIR RESEARCH AND DEVELOPMENT COMMAND, UNDER CONTRACT NO. AF19(604)1566.



# Raw data and notations

Let's denote a spatio-temporal process $S = (S(x,t); x \in \mathbb{R}^2, t \in \{t_1, \cdots, t_p\})$.

The temporal average of $S$ in denoted:

$$\bar{\mathbf{s}}^t(x) = \frac{1}{p} \sum_{k=1}^{p} S(x, t_k)$$

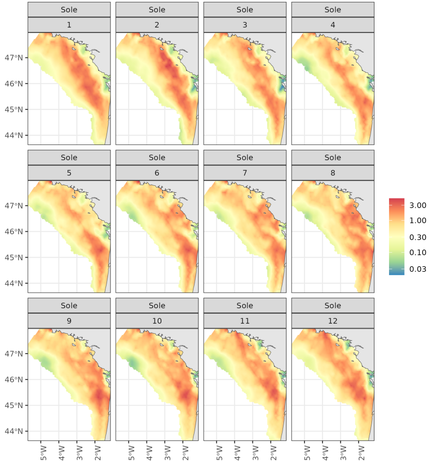
The time-centered space-time field:

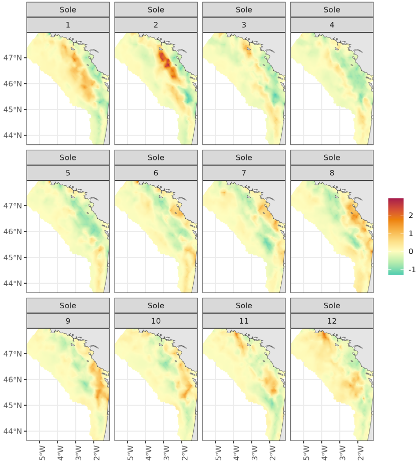
$$\mathbf{S}' = \left(\mathbf{s}_{t_1} - \bar{\mathbf{s}}^t, ..., \mathbf{s}_{t_p} - \bar{\mathbf{s}}^t\right).$$

Then, $\mathbf{S}'$ has the form:

$$\mathbf{S}' = \begin{pmatrix} S'(x_1, t_1) & S'(x_1, t_2) & \cdots & S'(x_1, t_p) \\ S'(x_2, t_1) & S'(x_2, t_2) & \cdots & S'(x_2, t_p) \\ \vdots & \ddots & \ddots & \vdots \\ S'(x_n, t_1) & S'(x_n, t_2) & \cdots & S'(x_n, t_p) \end{pmatrix}$$

# Raw data




(left) Montly spatial log-predictions log $S(x, t)$ of the hierarchical model. (right) Monthly anomalies of the spatial predictions $S^*(x, t)$. Each panel corresponds to the average distribution of prediction of anomalies for a month over the period 2008 - 2018.

# Basics of EOF

The spatio-temporal field is decomposed so that:

$$S'(x,t) = \sum_{m=1}^{r} p_m(x) \cdot \alpha_m(t) + \epsilon_m(x,t)$$

with $r$ the number of dimensions of the EOF ($r \leq min(n,p)$), $p_m(x)$ the spatial term of EOF and $\alpha_m(t)$ the temporal term of EOF for dimension $m$.
$\epsilon_m(x,t)$ is an error term.

Constraints:

- minimize $E = \sum_m \sum_x \sum_y \epsilon_m(x,t)$
- spatial terms and temporal terms are orthogonal

$$\langle p_i(\cdot); p_j(\cdot) \rangle = 0 \quad i \neq j$$

$$\langle \alpha_i(\cdot); \alpha_j(\cdot) \rangle = 0 \quad i \neq j$$

# Basics of EOF

This falls back to a diagonalisation issue through eigen-decomposition:

$$\mathbf{S}'\mathbf{S}'^T = \mathbf{C}_{\mathbf{S}'} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$$

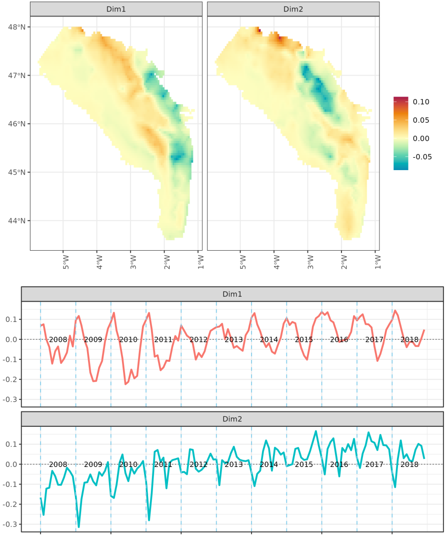
or through singular value decomposition:

$$\mathbf{S}' = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \quad \text{(SVD)}$$

- $\mathbf{C}_{\mathbf{S}'}$ is the covariance function of $\mathbf{S}'$
- $\mathbf{U}_{(n \times r)}$ contains the spatial factors $(p_m(x))$,
- $\mathbf{\Lambda}_{(r \times r)}$ contains the eigen values and $\mathbf{\Sigma}_{(r \times r)}$ contains the singular values of $\mathbf{S}'$.

  These quantifies the percentage of variance captured by each dimension.
  They are diagonal matrices with $\mathbf{\Lambda} = \mathbf{\Sigma}^2$
- $\mathbf{V}_{(p \times r)}$ contains the temporal loadings $(\alpha_m(t))$

# Illustration




(Top) Spatial factors for the two first dimensions of the EOF. (Bottom) Loadings for the two first dimensions of the EOF. Blue dashed vertical lines corresponds to the month of January for each year.

**Interpretation:**

$\mathbf{U}$ are spatial factors that capture the variance of $\mathbf{S}'$.

$\mathbf{V}$ are the temporal loadings that relate $\mathbf{S}'$ to the spatial factors in $\mathbf{U}$.

When the loading of dimensions $j$ denoted $\mathbf{V}_{j,\cdot}$ are high (resp. low) at time step $t$ then the process $\mathbf{s}'_t$ at this time step follows the spatial factor $\mathbf{U}_{j,\cdot}$ (resp. $-\mathbf{U}_{j,\cdot}$).

# Multivariate EOF

# Multivariate EOF

# Multivariate EOF

Let us denote by $k \in \{1, \cdots, s\}$ the number of variables

$S^{(k)}(x, t)$ is the value of the space time process for the location $x$, the time $t$ and the variable $k$.

To build the multivariate spatio-temporal matrix to be diagonalized, there are two options:

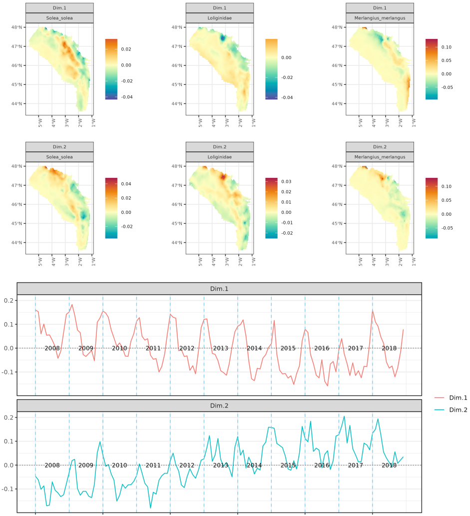
- binding the matrices by rows. The matrix is of dimension $(n \cdot s) \times p$

$$\mathbf{S}'^{(row)}_{multi} = \begin{pmatrix} \mathbf{S}'^{(1)} \\ \mathbf{S}'^{(2)} \\ \ldots \\ \mathbf{S}'^{(s)} \end{pmatrix}$$

- binding the matrix by columns. The matrix is of dimension $n \times (p \cdot s)$.

$$\mathbf{S}'^{(col)}_{multi} = \begin{pmatrix} \mathbf{S}'^{(1)} & ; & \mathbf{S}'^{(2)} & ; & \cdots & ; & \mathbf{S}'^{(s)} \end{pmatrix}$$

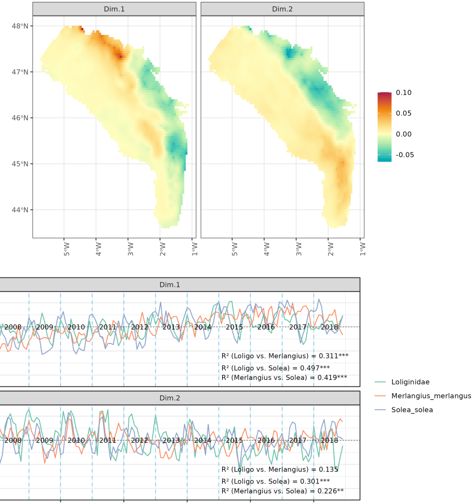
# Bind the matrices by rows




➠ **U** is of dimension $(n \cdot s) \times r$ and **V** is $p \times r$

Multivariate EOF on the matrix $\mathbf{S}'^{(row)}_{multi}$. (Top) Factor maps for each species and dimensions. (Bottom) Loadings for the two first dimensions.

# Bind the matrices by columns




➠ **U** is of dimension $n \times r$ and **V** is $(p \cdot s) \times r$

Multivariate EOF on the matrix $\mathbf{S}'^{(col)}_{multi}$. (Top) Factor maps for the two first dimensions. (Bottom) Loadings of each species for the two first dimensions.

# Constraining the EOF with an ancillary variable

# Constraining the EOF with an ancillary variable

**Aim:** "create new variables that are linear combinations of two (multivariate) data sets such that the correlations between these new variables are maximized" (Wickle et al., 2019).

Let's consider two spatio-temporal variables $S^{(1)}(x,t)$ and $S^{(2)}(x,t)$.

Now consider two new variables that are combinations of $S^{(1)}(x,t)$ and $S^{(2)}(x,t)$

$$a_k(t_j) = \sum_{i=1}^{n} \xi_{ik} S^{(1)}(x_i; t_j) = \boldsymbol{\xi}'_k \mathbf{s}^{(1)}_{t_j}$$

$$b_k(t_j) = \sum_{\ell=1}^{m} \psi_{\ell k} S^{(2)}(r_\ell; t_j) = \boldsymbol{\psi}'_k \mathbf{s}^{(1)}_{t_j}$$

The weights (*i.e.* the k$^t h$ canonical correlation) are the correlation between $a_k$ and $b_k$ with $k \in \{1, \cdots, \min\{n, m\}\}$:

$$r_k = \mathrm{corr}(\mathbf{a}_k, \mathbf{b}_k) = \frac{\mathrm{cov}(\mathbf{a}_k, \mathbf{b}_k)}{\sqrt{\mathrm{var}(\mathbf{a}_k)}\sqrt{\mathrm{var}(\mathbf{b}_k)}}$$

# Constraining the EOF with an ancillary variable

The correlation takes the form:

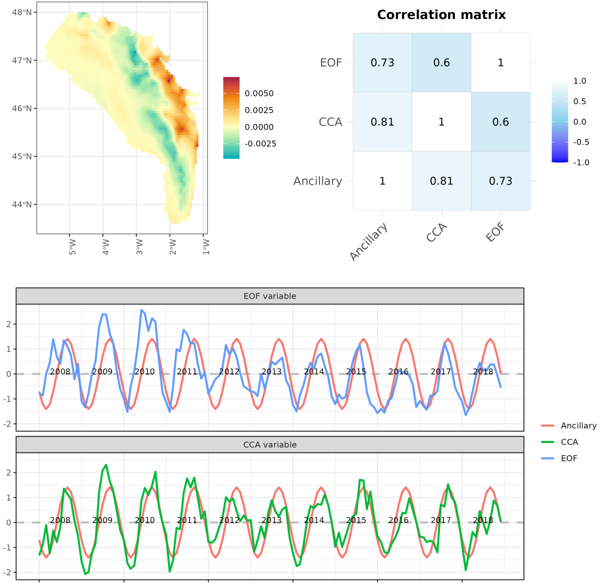
$$r_k = \frac{\boldsymbol{\xi}'_k \mathbf{C}_{S^1 S^2} \boldsymbol{\psi}_k}{(\boldsymbol{\xi}'_k \mathbf{C}_{S^1} \boldsymbol{\xi}_k)^{1/2} (\boldsymbol{\psi}'_k \mathbf{C}_{S^2} \boldsymbol{\psi}_k)^{1/2}}$$

- $\mathbf{C}_{S^1}$ and $\mathbf{C}_{S^2}$ are covariance matrices of dimension $m \times m$ and $n \times n$
- $\mathbf{C}_{S^1 S^2}$ is the covariance matrix between $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$ with dimension $m \times n$.
- The first pair of canonical variable corresponds to the weights $\boldsymbol{\xi}_1$ and $\boldsymbol{\psi}_1$ that maximize $r_1$.

**interpretation:**

The time series of the first few canonical variables ($\mathbf{a}_1$ and $\mathbf{b}_1$) typically match up fairly closely (they maximize correlation $r_1$)

The spatial patterns (the weights $\boldsymbol{\xi}_1$ and $\boldsymbol{\psi}_1$) show the areas in space that are most responsible for the high correlations.




Results of the canonical correlation analysis. (Top left) Canonical vectors $\boldsymbol{\xi}_1$ that maximise correlation between $\mathbf{a}_k$ and $\mathbf{b}_k$. (Top right) Correlation matrix between the first time series of the EOF $U_{\cdot,1}$, the EOF time series in the CCA $\mathbf{a}_1$ and the ancillary variable. (Bottom) Comparison of the EOF variables with the ancillary variable and the CCA variables with the ancillary variable. These time series are standardized.

# Are EOF truly orthogonal?

# Are EOF truly orthogonal?

EOFs are statistically orthogonal *i.e.* for $i \neq j, \{\mathbf{u}_i, \mathbf{u}_j\} = 0$

Bu they are not spatially orthogonal *i.e.* the cross-covariance of the different EOF maps is not necessarily 0.

EOM consist in a two step analysis:

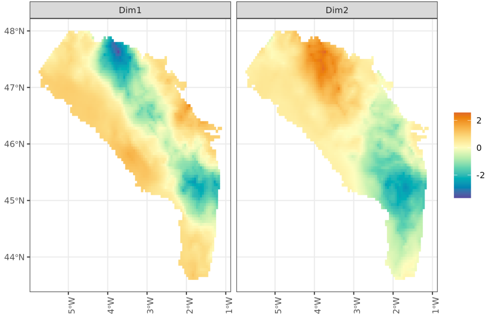
- perform an EOF on $\mathbf{S}'$ $\rightarrow$ obtain statistically decorrelated plans
- compute the variogram and the cross-variogram of $\mathbf{U}$ for a specific ditance $r$ that we denote $\mathbf{\Gamma}_r$ of dimension $(p \times p)$.

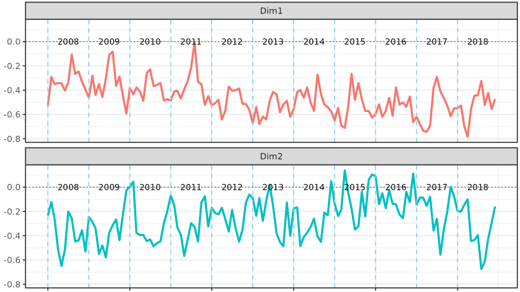
$$\mathbf{\Gamma}_r = \boldsymbol{U}_{\text{eom}} \mathbf{\Sigma}_{eom} \mathbf{V}_{\text{eom}}^T$$

Plans are ordered by increasing variance explained in $\mathbf{S}'$

$$p_k = \frac{\text{tr}\left(C_{\hat{\mathbf{S}}_k}\right)}{\text{tr}\left(C_{\mathbf{S}}\right)}$$

where $p_k$ is the proportion of variance explained by factor $k$ and $\hat{\mathbf{S}}_k$ is the projection of $\mathbf{S}$ in the space of the EOM.




(Top) Spatial factors obtained by EOM for the two first dimensions. (Bottom) Temporal loadings for the first two dimensions.

# Conclusion

# Conclusion: some bibliographic metrics

| sources | articles |
|---|---|
| JOURNAL OF CLIMATE | 391 |
| CLIMATE DYNAMICS | 271 |
| JOURNAL OF GEOPHYSICAL RESEARCH: OCEANS | 257 |
| INTERNATIONAL JOURNAL OF CLIMATOLOGY | 256 |
| GEOPHYSICAL RESEARCH LETTERS | 142 |
| JOURNAL OF THE ATMOSPHERIC SCIENCES | 126 |
| MONTHLY WEATHER REVIEW | 103 |
| REMOTE SENSING | 99 |
| JOURNAL OF GEOPHYSICAL RESEARCH ATMOSPHERES | 97 |
| THEORETICAL AND APPLIED CLIMATOLOGY | 85 |
| PROCEEDINGS OF SPIE - THE INTERNATIONAL SOCIETY FOR OPTICAL ENGINEERING | 75 |
| ATMOSPHERE | 72 |
| CONTINENTAL SHELF RESEARCH | 72 |
| JOURNAL OF PHYSICAL OCEANOGRAPHY | 72 |
| QUARTERLY JOURNAL OF THE ROYAL METEOROLOGICAL SOCIETY | 68 |

# Infering loading factors from sparse data

- EOF are mainly applied in climate science. . .
- but progressively being transferred to other fields of applications

➠ In **ecology**, EOF must be computed from sparse data (`VAST`, Jim Thorson)

For instance, we could consider a **hierarchical model** where the **latent field** take the form:

$$log(S(x,t)) = \beta(t) + \omega(x) + \sum_{f=1}^{N_f} \lambda(t,f)\epsilon(x,f)$$

where $f$ are the dimension of the EOF with $N_f$ dimensions, $\boldsymbol{\omega} \sim \mathcal{MG}(O, \boldsymbol{\Sigma}_\omega)$ and $\boldsymbol{\epsilon} \sim \mathcal{MG}(O, \boldsymbol{\Sigma}_\epsilon)$.
$\lambda(t,f)$ are the loadings and $\epsilon(x,g)$ are the EOF.

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

and **the observations** are zero-inflated and takes the form:

$$\Pr(y_i = Y) = \begin{cases} 1 - p_i & \text{if } y_i = 0 \\ p_i \times \mathcal{L}(y_i; \log S(x_i, t_i); \sigma^2) & \text{if } y_i > 0 \end{cases}$$

Where $Y$ is an observations as random variable and $y_i$ is the realized observation. $p_i$ is the probability to obtain a positive observation. $\mathcal{L}$ is the probability of the positive observations.

# Take home message

- EOF and EOM provide patterns that capture variance of a spatio-temporal dataset while being orthogonal.
- EOF can be extended to several variables.
- EOF can be constrained with an ancillary variable through CCA.
- EOF maps are only statistically orthogonal.
- EOM maps are spatially orthogonal ➠ loadings are also harder to interpret.
- EOM could be realized on time steps, not on locations → but not both at the same time.

**Open questions:**

- How to perform dimension-reduction and take into account both spatial and temporal correlation?
- How to perform dimension reduction without the need if svd?

# Bibliography


Bez, N., Renard, D., & Ahmed-Babou, D. (2023). Empirical Orthogonal Maps (EOM) and Principal Spatial Patterns: Illustration for Octopus Distribution Off Mauritania Over the Period 1987–2017. Mathematical Geosciences, 55(1), 113-128.

Dufresne, J. L., Foujols, M. A., Denvil, S., Caubel, A., Marti, O., Aumont, O., . . . & Vuichard, N. (2013). Climate change projections using the IPSL-CM5 Earth System Model: from CMIP3 to CMIP5. Climate dynamics, 40, 2123-2165.

Garrigues, S., Allard, D., & Baret, F. (2008). Modeling temporal changes in surface spatial heterogeneity over an agricultural site. Remote Sensing of Environment, 112(2), 588-602.

Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., & Kelling, S. (2009). eBird: A citizen-based bird observation network in the biological sciences. Biological conservation, 142(10), 2282-2292.

Thorson, J. T., Cheng, W., Hermann, A. J., Ianelli, J. N., Litzow, M. A., O'Leary, C. A., & Thompson, G. G. (2020). Empirical orthogonal function regression: Linking population biology to spatial varying environmental conditions using climate projections. Global Change Biology, 26(8), 4638-4649.

Wikle, Christopher K., Andrew Zammit-Mangion, and Noel Cressie. Spatio-temporal statistics with R. Chapman and Hall/CRC, 2019.