Effective probability distribution approximation for non-stationary non Gaussian random fields - An application to precipitation

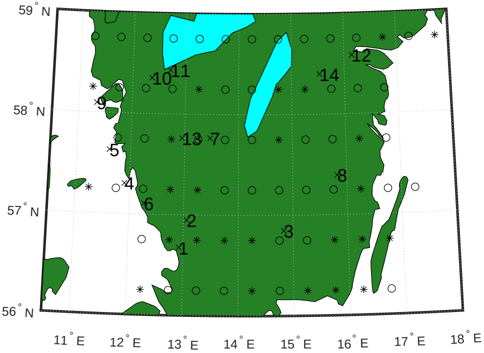
Anastassia Baxevani,
Joint work with D. Hristopulos and C. Andreou

University of Cyprus

SWGEN 2025
Grenoble, France

Contents

Motivation



A few issues in stochastic modeling in space (and time)

- Marginal distribution, Dependence structure, (Dynamics)
- Computational cost

Gaussian based Models

Usual Method of Choice:

- Gaussian based model (mean + covariance structure) works for Gaussian and non-Gaussian data:
- Gaussian assumption is usually a working hypothesis
- Non Gaussian data - Gaussian model + marginal transformation (Gaussian anamorphosis)
- Pros: Simple to use, explicit expressions, closed under linear transformations and conditioning. Can generate mass at zero by truncation.
- Cons: Transformation operates on marginal distribution, difficult to see what happens to joint densities, computational complexity

Alternative Models

- Few in closed-form models: Wishart, gamma, t-Student, Laplace
- Integrals (e.g. moving averages) with non-Gaussian noise, e.g. heavy-tailed Laplace random field moving averages

Even if closed-form models are possible computational complexity of processing joint pdf's for large data is really prohibited.

Effective distribution model (EDM)

The core idea of EDM, is that we can simulate realizations of missing values at prediction sites $\tilde{\mathrm{s}}_p \in \mathcal{P}$, conditionally on the data $\mathrm{x} := \mathrm{x}(\mathrm{s})$, while preserving the spatial correlations with the nearby locations, using the univariate pdf:

$$f_{\mathrm{eff}}\left(y_p; \boldsymbol{\psi}(\tilde{\mathrm{s}}_p; \boldsymbol{\psi}_1, \ldots \boldsymbol{\psi}_N)\right) \; .$$

Schematic Description of the methodology

- Choose the functional form of $f_{\text{eff}}(\cdot)$ - univariate effective pdf - based on either empirical knowledge or from fitting the sample data.
- Fit the model to the data at each one of the N sample sites in the set $\mathcal{S}$ - which produces the parameters vectors $\boldsymbol{\psi_1, \ldots \psi_N}$.
- Predict the value of the parameter vector at the prediction site $\tilde{\mathrm{s}}_p$:

$$\boldsymbol{\psi}_p^* = \boldsymbol{\psi}(\tilde{\mathrm{s}}_p; \boldsymbol{\psi}_1, \ldots \boldsymbol{\psi}_N)$$

  - using stochastic methods - like krigging
  - using deterministic methods - like kernel regression
- Simulate y_p from the conditional pdf $f_{\text{eff}}\left(y_p; \boldsymbol{\psi}(\tilde{\mathrm{s}}_p; \boldsymbol{\psi}_1, \ldots \boldsymbol{\psi}_N)\right)$, using a simulation method that further imposes spatial correlations between the prediction site and its neighbors.

Simulation algorithm - basic ideas

- The conditional pdfs $f_{\text{eff}}(y_p, \psi_p^*)$ incorporate the local spatial variation of ψ_p^*.
- However, this does not ensure spatial continuity of the reconstructed precipitation field at neighboring locations.
- We propose two simulation algorithms with this intend.
- Spatial correlations are imposed by selecting the level of the effective cdf at target sites based on probability levels at neighboring sampling sites.

Simulation algorithm 1: "Frozen" Sample (FS)

- For $\tilde{\mathrm{s}}_p \in \mathcal{P}$ (Prediction set), define a bounded region $\mathcal{B}(\tilde{\mathrm{s}}_p)$ which includes the n_b nearest neighbors of $\tilde{\mathrm{s}}_p$ that lie in $\mathcal{S}$ (default $n_b = 5$).
- Randomly select one element from $\mathcal{B}(\tilde{\mathrm{s}}_p)$ that corresponds, say, to the sampling location s_k, where $k \in \{1, \ldots, N\}$.
- Determine the probability level at location s_k by means of $p(\mathrm{s}_k) = F_{\mathrm{eff}}(x_k; \hat{\boldsymbol{\psi}}_k)$, where $\hat{\boldsymbol{\psi}}_k = \boldsymbol{\psi}(\mathrm{s}_k)$, and F_{eff} is the cdf of f_{eff}.
- Assign the probability level $p(\mathrm{s}_k)$ to the location $\tilde{\mathrm{s}}_p$, i.e., $p = p(\mathrm{s}_k)$.
- Assign to the grid location $\tilde{\mathrm{s}}_p$ the value $y_p = F_{\mathrm{eff}}^{-1}(p; \hat{\boldsymbol{\psi}}_p)$.
- Repeat for all prediction points.

Sequential Updating Algorithm

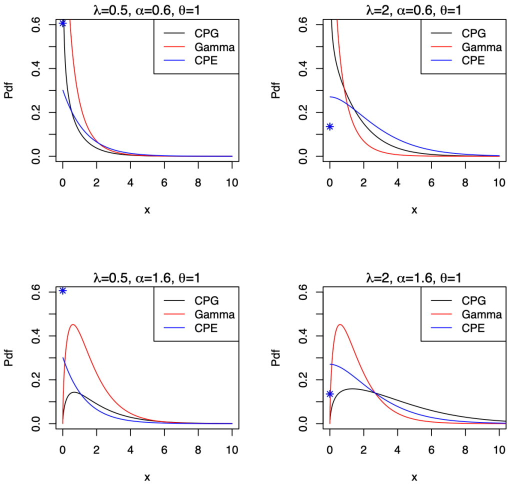
The Sequential Updating (SU) algorithm uses a continuously updated “sample set” which incorporates the prediction sites where the algorithm has already assigned values.

Compound Poisson gamma (CPG) distribution

$$X = \sum_{i=1}^{N_c} \Gamma_i, \quad N_c \sim \mathrm{Poisson}(\lambda), \quad \Gamma_i \sim \mathrm{iid\ gamma}(\alpha, \theta)$$

- Mixed type with an atom at zero $\lambda : \mathbb{P}(X = 0) = e^{-\lambda} = \mathbb{P}(N_c = 0)$ - dry conditions;
- CPG belongs to the family of Tweedie distributions; estimation is by mle (numerically) and is implemented in the `Tweedie` package in R.

CPG density plots



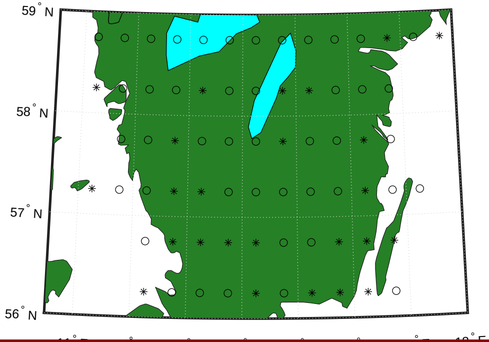
CPG vs gamma

We prefer CPG because of :

- mixed type distribution. The zeros that correspond to dry conditions are produced naturally;
- the total precipitation amount during a day, is generated as a sum of precipitation amounts during N_c individual rain events Γ_i which in principle, and depending on the resolution of the available data, can have different shape and rate;
- has in general fatter tails than the gamma distribution.

Reanalysis daily precipitation data - South Sweden

https://downloads.psl.noaa.gov/Datasets/cpc_global_precip/
Climate Prediction Center (CPC) at 70 nodes with spatial resolution $\approx 0.5^o \approx 55.65$ km) from 1/1/1979 -31/12/2019

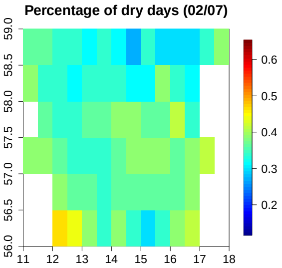


Precipitation model - application

- Our approach leads to a spatial precipitation model indexed by day.
- Each point in the sampling set is assigned a pdf which represents the daily precipitation for that point for the specific day of the year.
- Distribution is estimated by the precipitation records for the specific day over the entire period of observation.
- Estimated parameters are location dependent.
- No need for temporal or spatial stationarity assumption.

Some Statistics for 2nd July

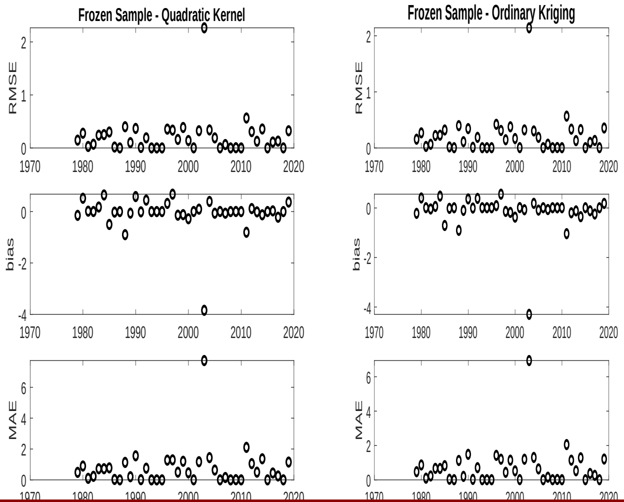
Mean	StD	Skewness	Kurtosis	Maximum
1.9 mm	4.2 mm	6.1	55.9	55.7 mm



Simulation Scheme

1. At locations with data, denoted by " $\circ$ " estimate $\psi(\mathrm{s})$ for July 2 using the precipitation records for this day over all available years.
2. Predict ψ_p^* at the remaining 25 locations denoted by " $\star$ " locations
   - kernel regression with quadratic and Gaussian kernel
   - krigging equations with Matérn covariance
3. Using EDM we generate 100 realizations conditionally on the precipitation values at the sampling locations for each year (i.e., 40×100 simulations per prediction location).
4. The CPG-EDM prediction at each site for a specific year is given by the median over the 100 realizations that correspond to that year.
5. We assess the performance of the CPG-EDM approach by means of the validation scores.

Continuous scores

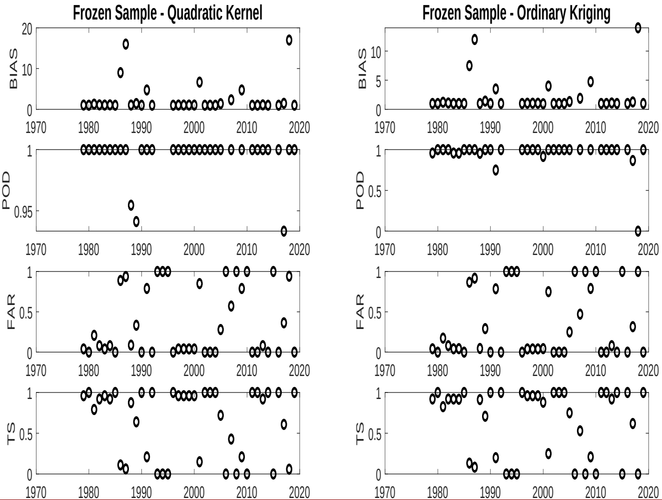


Categoricals scores

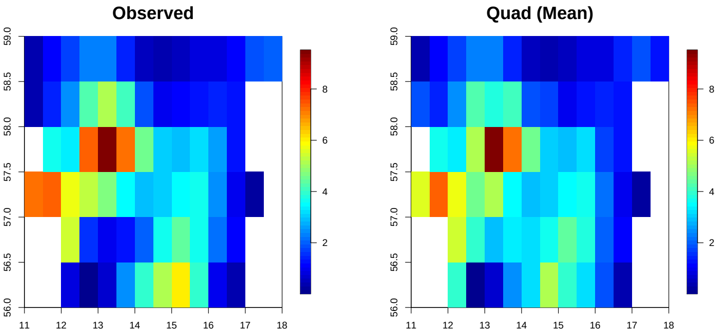
Categorical scores		
Index	Description	Formula
BIAS	Bias score	$\frac{\text{hits}+\text{false alarms}}{\text{hits}+\text{misses}}$
POD	Probability of detection	$\frac{\text{hits}}{\text{hits}+\text{misses}}$
FAR	False alarm ratio	$\frac{\text{false alarms}}{\text{hits}+\text{false alarms}}$
TS	Threat score	$\frac{\text{hits}}{\text{hits}+\text{misses}+\text{false alarms}}$

Table: Descriptions and definitions for categorical scores. Hits refers to the number of cases the predictions matched the observations; false alarms refers to the number of false precipitation predictions; misses counts the failures to predict a precipitation event.

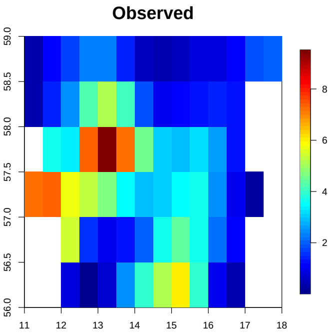
Categoricals scores

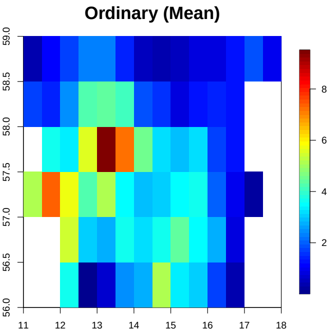


Field reconstruction



Field reconstruction





Field reconstruction - statistics

	MAE	MSE	RMSE	MAPE	Pearson	R-squared
Quadratic	1.063	1.85	1.36	76.65	0.754	0.568
Ordinary	1.091	1.771	1.331	74.698	0.764	0.583

Table: Comparison of the generated precipitation amounts using quadratic kernel and ordinary krigging compared to the baseline 2 July 2014.

Conclusions-Remarks

- The EDM approach decomposes the joint problem to local densities and thus is suitable for large data sets and non-Gaussian and non-stationary data.
- By coupling the effective pdf method with computationally efficient conditional simulation algorithms, we obtained promising results in reconstructing spatial data gaps in sets with complex dependencies (examples not shown here).
- The CPG distribution allows modeling intermittence and consider multiple rain events per day.
- The EDM-based algorithms were used for the reconstruction of spatial data: