Hidden Markov Models for precipitation

Pierre Ailliot
Université de Brest

Joint work with
Peter Thomson
Statistics Research Associates (NZ)

Context

- **Part of the project**

  "Climate-related risks for energy supply and demand"

- **Important part is to assess future climate variability**
  - **Based on past observations…**
  - **…and expected behavior of ENSO, IPO, human-induced climate change,...**
- **Aim : developing a stochastic models to simulate realistic scenarios for the future climate. First step:**
  - **Focus on rainfall**
  - **Reproduce past conditions**
- **Simulated rainfall used as input in models of large-scale hydrology and river flows**
  - **Estimate risks related to production of hydroelectricity (60% of NZ electricity)**

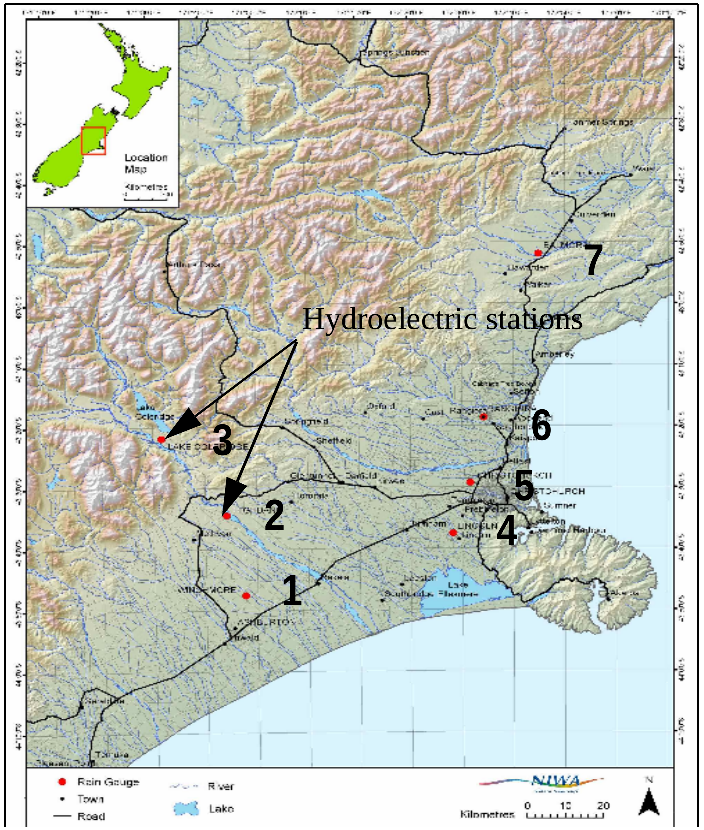
Outline

- Data
- Basic Hidden Markov Model (HMM)
- HMM with truncated Gaussian distributions
- Conclusion

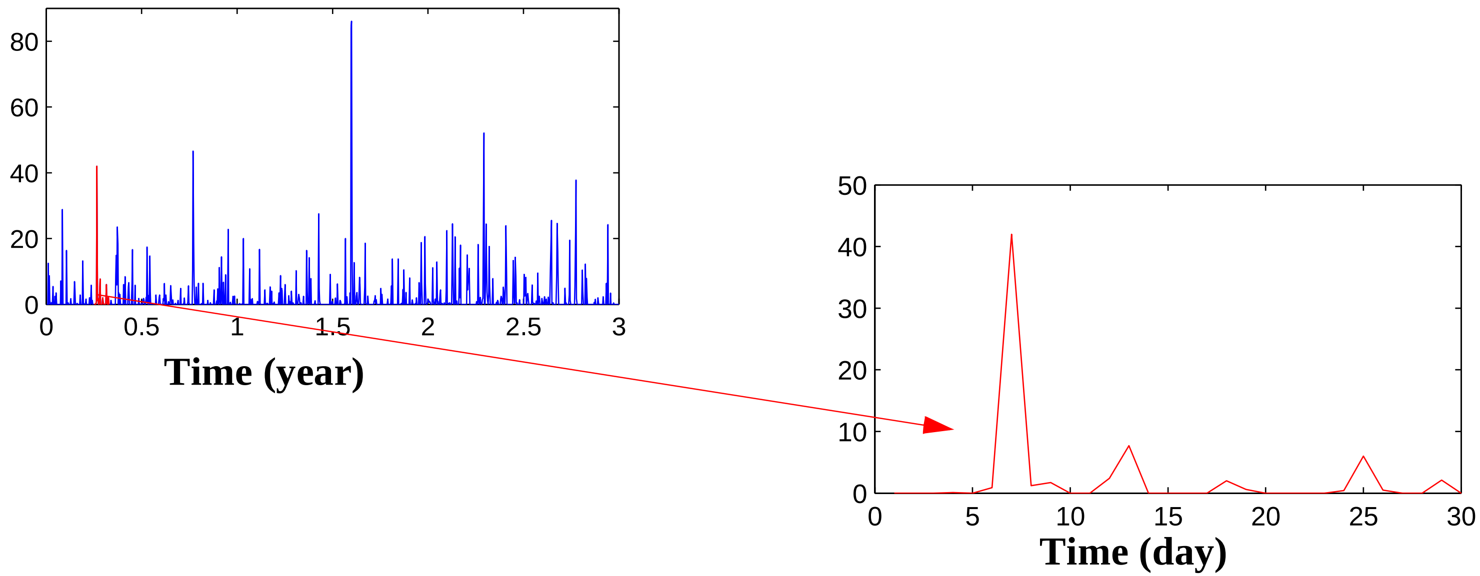
Data

Data

- **Rainfall data in New Zealand**
  - **Daily rainfall**
  - **K=7 locations**
- $\boldsymbol{Y_t = (Y_t(1), \ldots, Y_t(K))'}$
  - $\boldsymbol{Y_t(k) \in R^+}$ **: rainfall (mm) during day t at location k**
- **26 years**
  - **Focus on April**



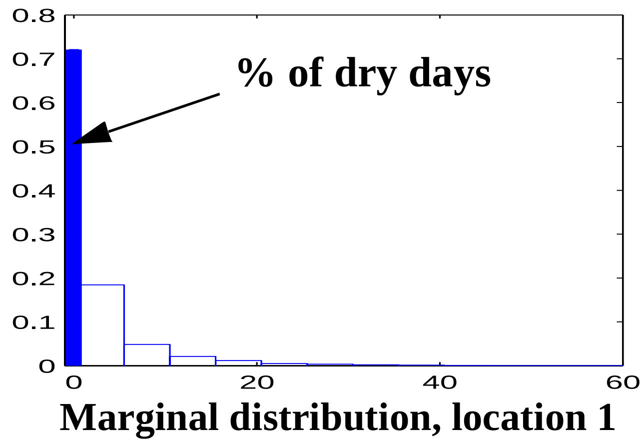
- **Time structure (location 1)**



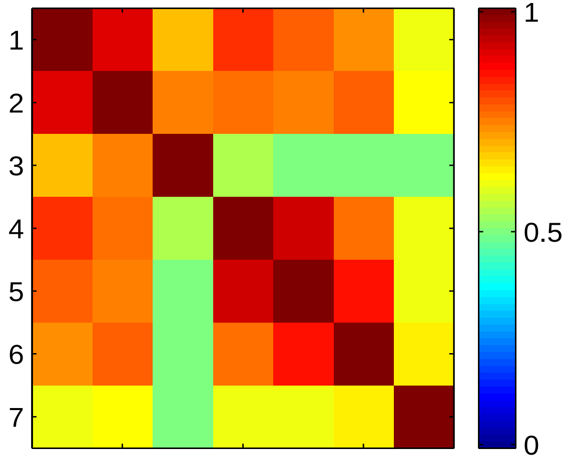
- **Marginal distribution**
  - **Two components:**

    . $Y_t(k) \;=\; 0$ **if no rainfall occurs**

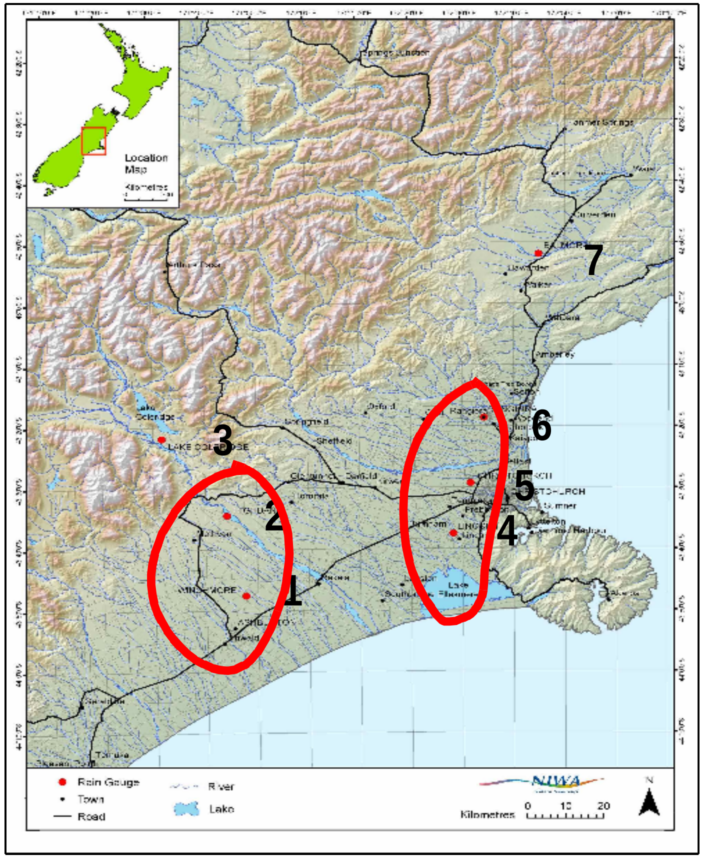
    . $Y_t(k) > 0$ **if a rainfall occurs**



- **Spatial structure**

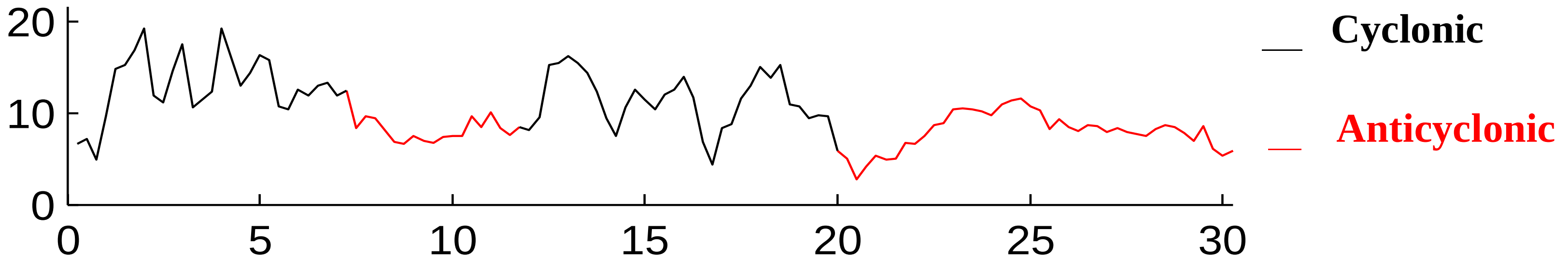


corr(Y_t)



Basic HMM: model description (Zucchini & Guttorp (1991))

- **Existence of "weather types" in meteorological time series**
  - **Another meteorological time series**



Wind speed (in m/s), Brittany, January 1991

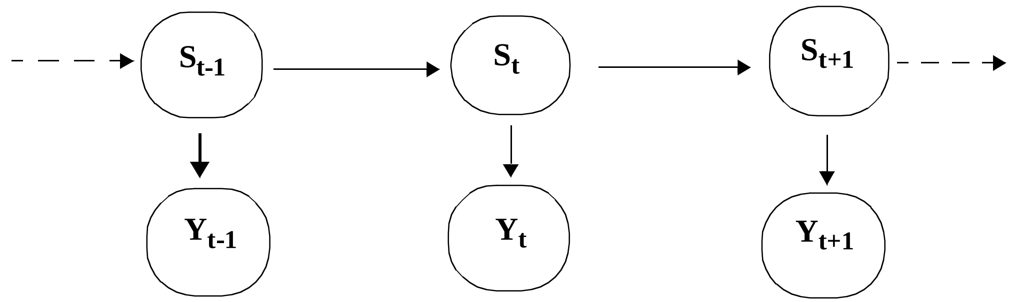
  - **The weather type depends on the position of high pressure systems, frontal systems...**
  - **Introduced as a hidden process** $S_t \in \{1 \ldots Q\}$

<u>Basic HMM: model description</u> (Zucchini & Guttorp (1991))

- **Existence of "weather types" in meteorological time series**
  - **Introduced as a hidden process** $S_t \in \{1 \ldots Q\}$
- **<u>Time structure:</u> HMM**
  - $P(S_t | S_{t-1} = s_{t-1}, Y_{t-1} = y_{t-1}, \ldots, S_0 = s_0, Y_0 = y_0) = P(S_t | S_{t-1} = s_{t-1})$
  - $P(Y_t | S_t = s_t, S_{t-1} = s_{t-1}, Y_{t-1} = y_{t-1}, \ldots, S_0 = s_0, Y_0 = y_0) = P(Y_t | S_t = s_t)$

<u>Weather type</u> (not observed)

<u>Precipitation</u> (observed)



  - **...Dynamics induced only by** $\{S_t\}$**!**

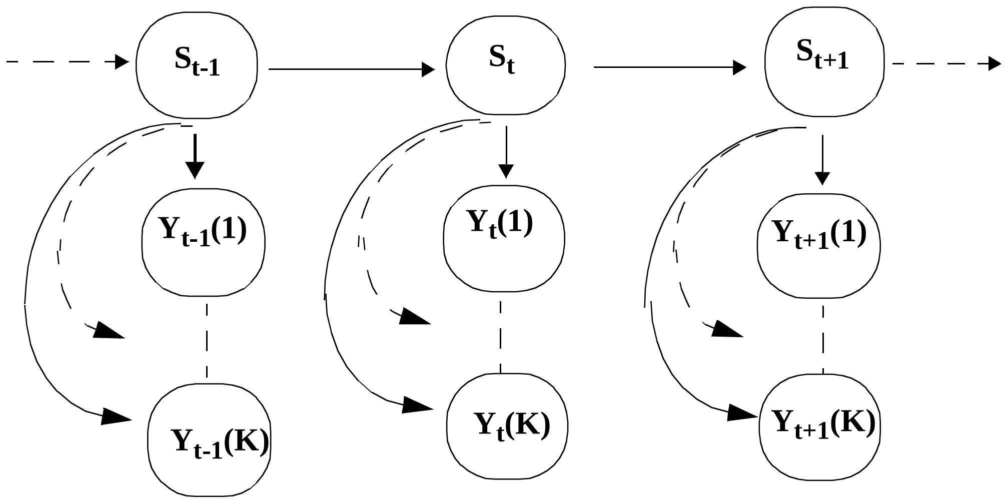
Basic HMM: model description (Zucchini & Guttorp (1991))

- **Existence of "weather types" in meteorological time series**
  - **Introduced as a hidden process** $S_t \in \{1 \dots Q\}$
- **Time structure: HMM**
  - **...Dynamics induced only by** $\{S_t\}$**!**
- **Spatial structure: conditional independence**

$$p(Y_t(1), \dots Y_t(K) | S_t = s_t) = \prod_{k \in \{1 \dots K\}} p(Y_t(k) | S_t = s_t)$$

Weather type (not observed)

Precipitation (observed)



- **Spatial dependence induced only by** $\{S_t\}$**!**

Basic HMM: model description (Zucchini & Guttorp (1991))

- **Existence of "weather types" in meteorological time series**
  - **Introduced as a hidden process** $S_t \in \{1 \ldots Q\}$
- **Time structure: HMM**
  - **...Dynamics induced only by** $\{S_t\}$**!**
- **Spatial structure: conditional independence**

$$p(Y_t(1), \ldots Y_t(K) | S_t = s_t) = \prod_{k \in \{1 \ldots K\}} p(Y_t(k) | S_t = s_t)$$

  - **Spatial dependence induced only by** $\{S_t\}$**!**

$$p(Y_t(k) \in dy | S_t = s) = \begin{cases} 1 - p_k^{(s)} & \text{if } 0 \in dy \\ p_k^{(s)} \gamma(y; \alpha_k^{(s)}, \beta_k^{(s)}) & \text{if } 0 \notin dy \end{cases}$$

  - $\gamma(y; \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} y^{\alpha - 1} exp\left(-\frac{y}{\beta}\right)$
  - $p_k^{(s)} \in [0, 1], \alpha_k^{(s)} > 0, \beta_k^{(s)} > 0$

Parameter estimation

- $3KQ + Q^2 - 1$ parameters
- EM algorithm to compute the Maximum Likelihood Estimates
  - Several random starting values to avoid local extrema
- Model selection
  - First selection with AIC and BIC

Q	1	2	3	4	5	6
AIC	17404	14317	13436	13213	13144	**12990**
BIC	17502	14523	13760	**13663**	13731	13722

- Focus on the model with $Q = 4$
  - Easier interpretation than models with $Q \geq 5$
  - Produce "more realistic" synthetic time series than models with $Q \leq 3$

Model validation

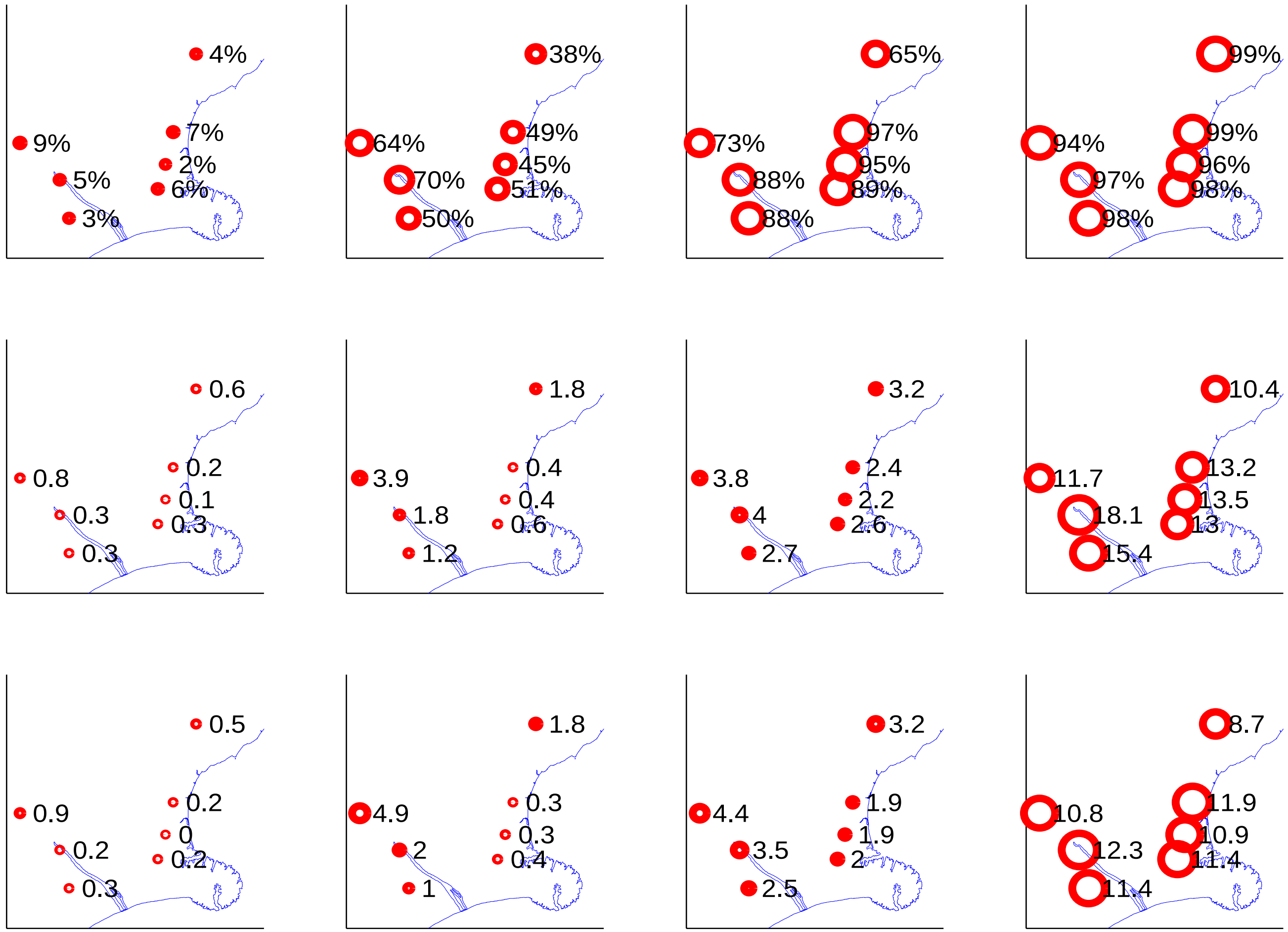
- **Meteorological interpretability**
  - **Emission probabilities**

Regime 1 **Regime 2** **Regime 3** **Regime 4**

$P(Y_t>0|S_t)$

$E[Y_t|S_t,Y_t>0]$

$var(Y_t|S_t,Y_t>0)$



Model validation

- **Meteorological interpretability**
  - **Emission probabilities**
  - **Transition matrix, stationary distribution, mean durations**

0.70	0.15	0.09	0.05
0.49	0.18	0.20	0.12
0.35	0.31	0.17	0.16
0.21	0.29	0.25	0.25

0.56
0.20
0.14
0.10

3.33
1.22
1.20
1.33

  - **Summary**

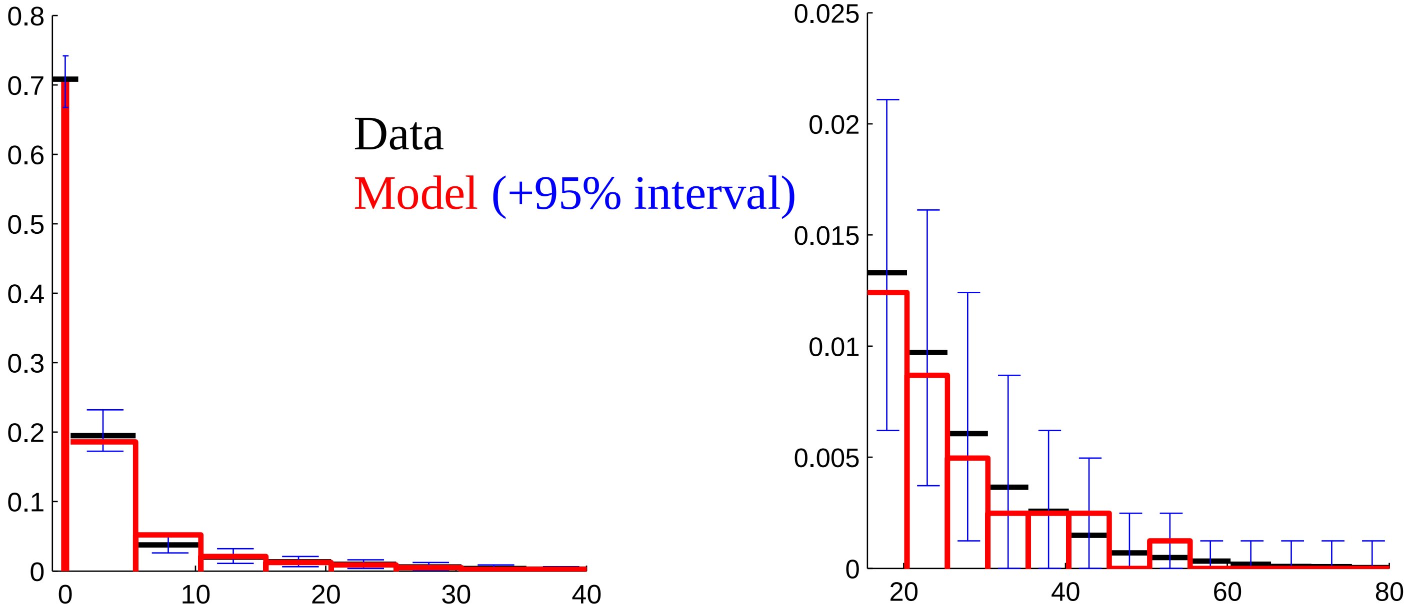
    Regime 1: **dry conditions, "long" persistence**

    Regime 2 and 3: **intermediate patterns, regional differences, higher rainfall in regime 3, short persistence**

    Regime 4: **heavy rainfall**
  - **Similar meteorological interpretation for other datasets**

Model validation

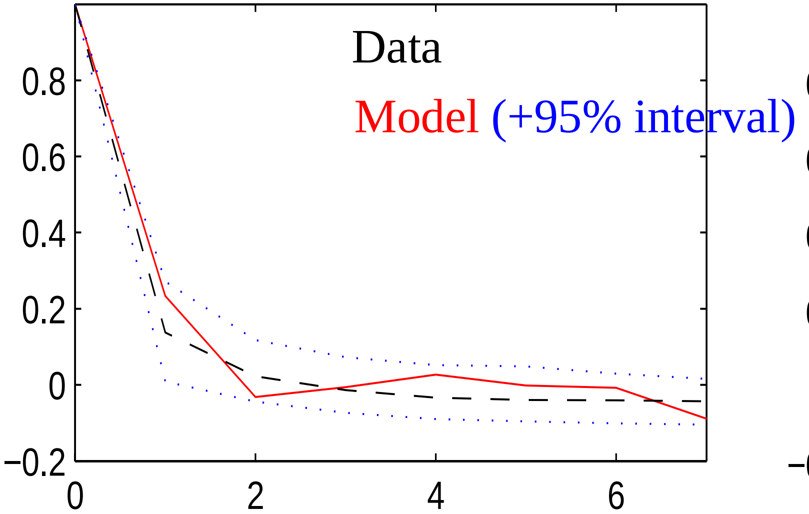
- **Meteorological interpretability: OK**
- **Realism of artificial sequences simulated with the model**
  - **Marginal distribution**



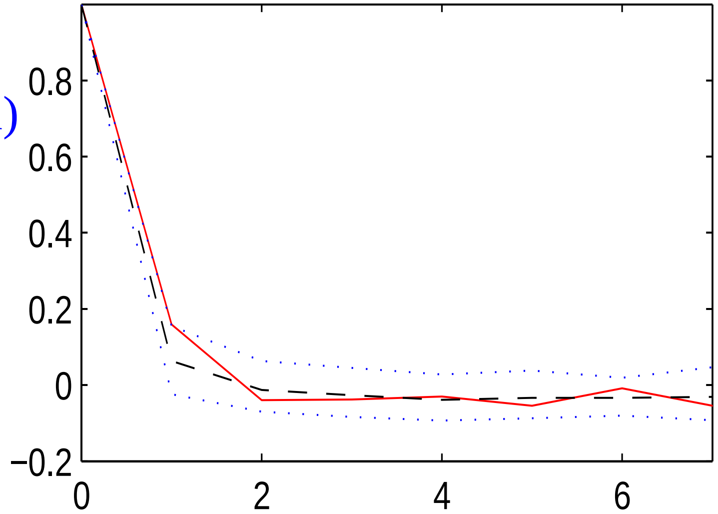
Location 1 (Winchmore)

Model validation

- **Meteorological interpretability: OK**
- **Realism of artificial sequences simulated with the model**
    - **Marginal distribution: OK**
    - **Dynamics at the different locations**



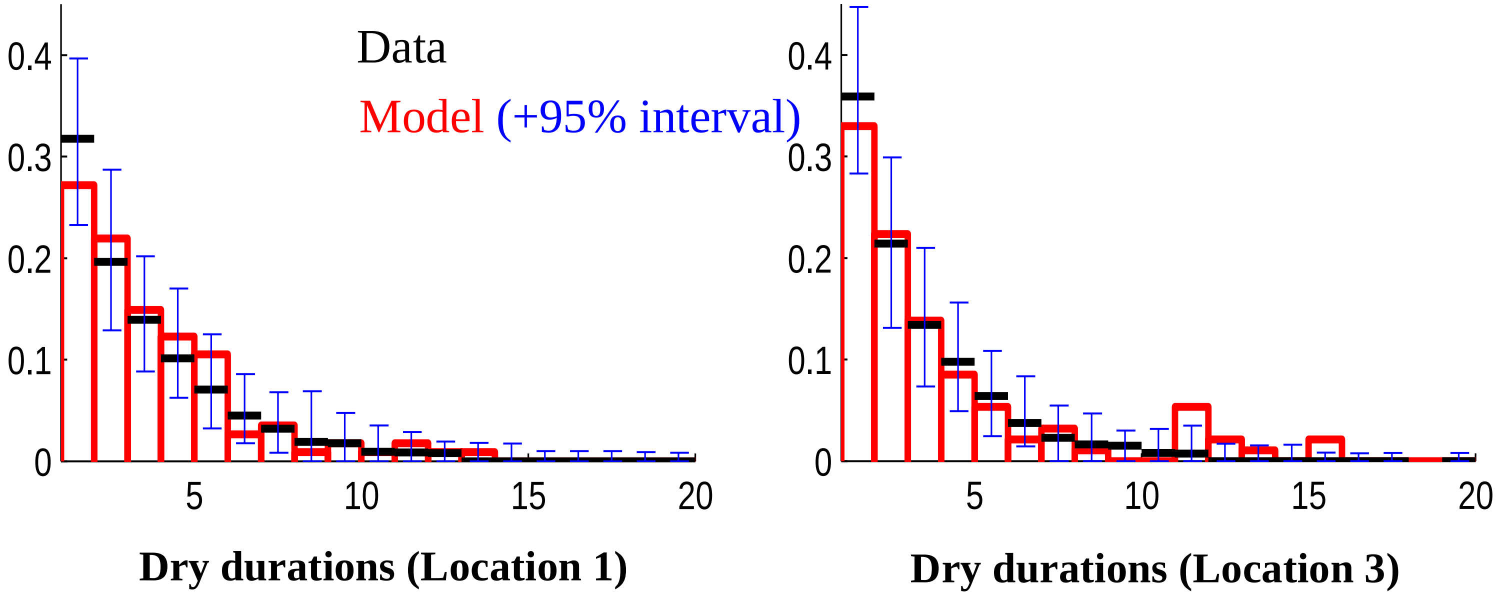
Autocorrelation (Location 1)



Autocorrelation (Location 3)

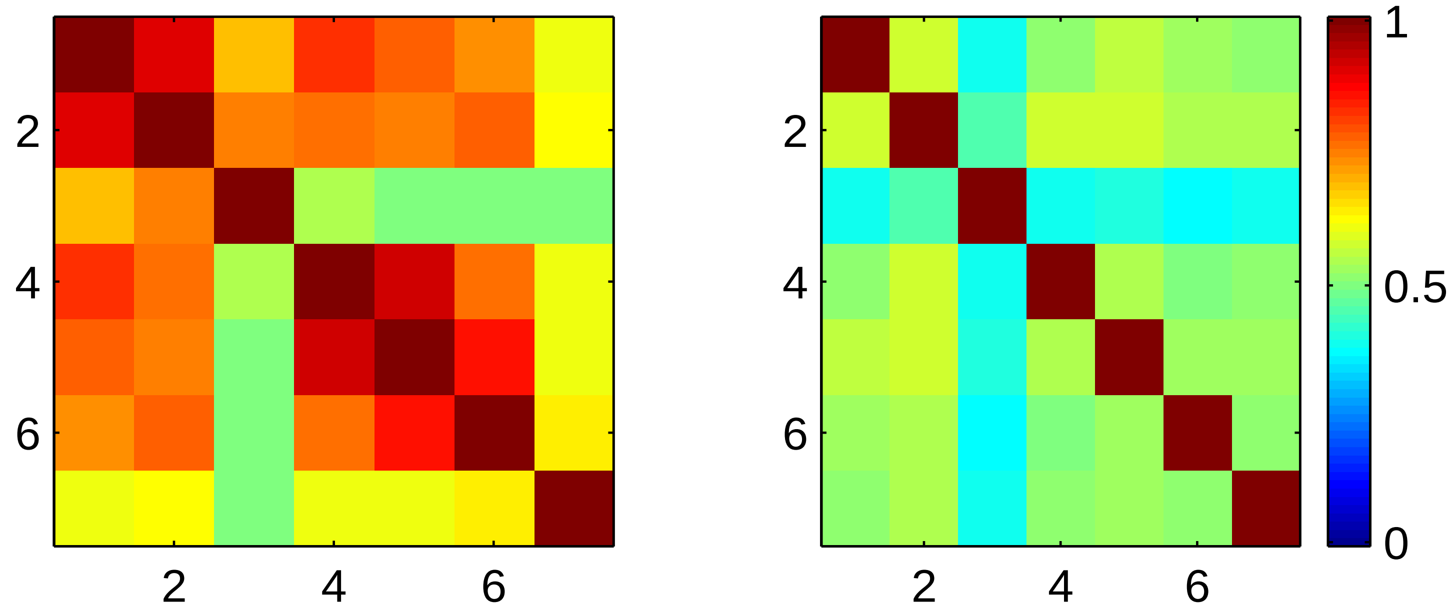
Model validation

- **Meteorological interpretability: OK**
- **Realism of artificial sequences simulated with the model**
  - **Marginal distribution: OK**
  - **Dynamics at the different locations**



Model validation

- Meteorological interpretability: OK
- Realism of artificial sequences simulated with the model
  - Marginal distribution: OK
  - Dynamics at the different locations: ~OK
  - Spatial structure

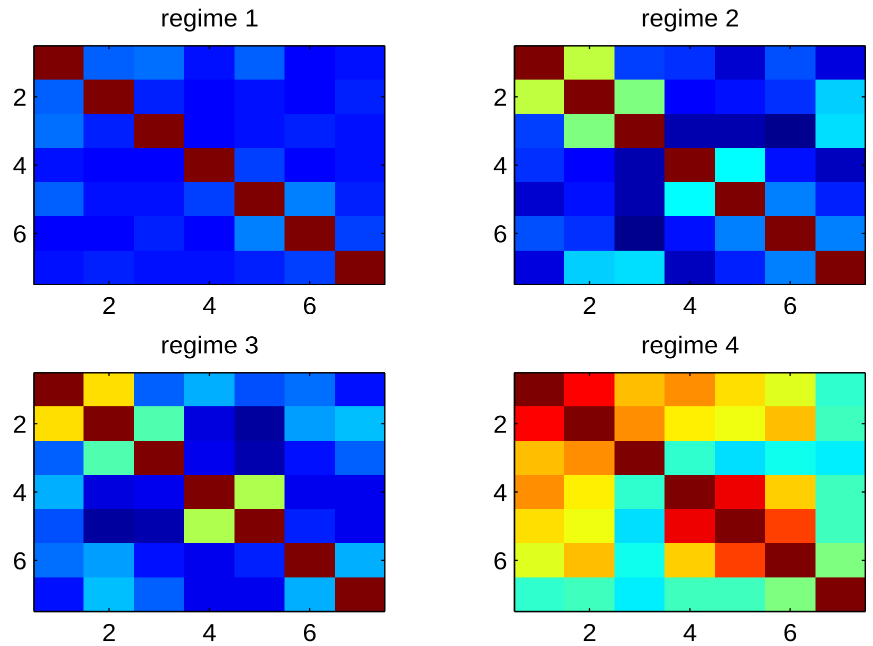


Data **Model**

Model validation

- **Meteorological interpretability: OK**
- **Realism of artificial sequences simulated with the model**
  - **Marginal distribution: OK**
  - **Dynamics at the different locations: ~OK**
  - **Spatial structure: not good!**
- **…Need for a better model!**
  - **Existence of residual spatial structure within the weather types**

Empirical correlation matrices in the different weather types (identified via the Viterbi algo.)



Model validation

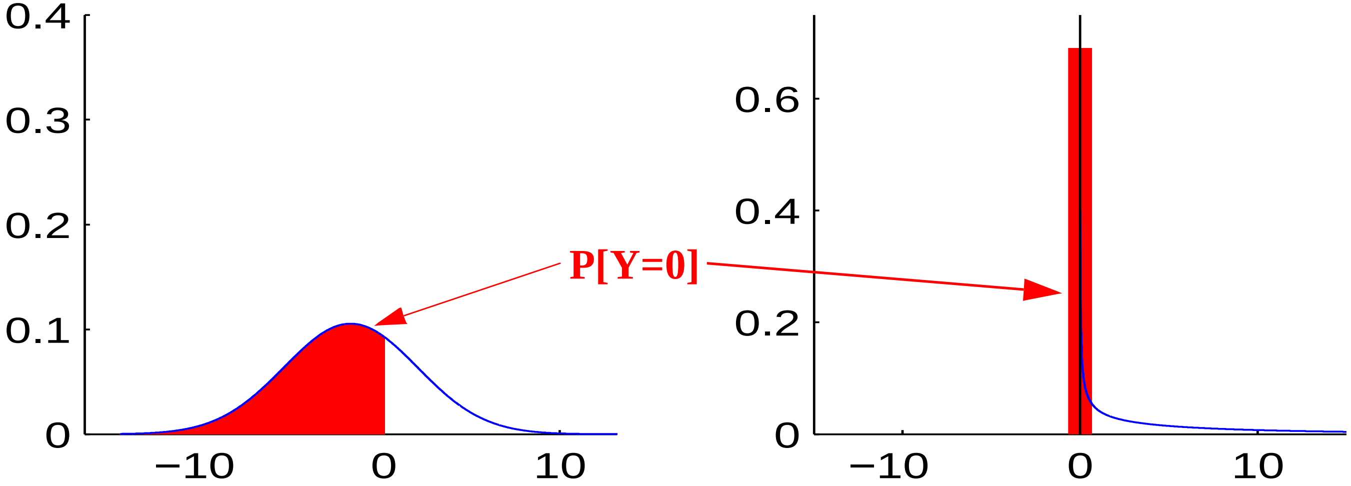
- **Meteorological interpretability: OK**
- **Realism of artificial sequences simulated with the model**
    - **Marginal distribution: OK**
    - **Dynamics at the different locations: ~OK**
    - **Spatial structure: not good!**
- **…Need for a better model!**
    - **Existence of residual spatial structure within the weather types**
- **Introduce spatial structure in the emission probabilities**
    - **Need model for multivariate mixed discrete-continuous distributions**
    - **...Truncated Gaussian random fields**

HMM with truncated Gaussian fields: model description

- **If $S_t = s$ then**

$$Y_t(k) = \begin{cases} 0 & \text{if } W_t(k) \le 0 \\ W_t(k)^{\beta^{(s)}(k)} & \text{if } W_t(k) > 0 \end{cases}$$

  - $W_t = m^{(s)} + H^{(s)} Z_t$ **with** $Z_t \sim N(O, I)$ **i.i.d**
  - $m^{(s)} \in R^K$, $\Sigma^{(s)} = H^{(s)}(H^{(s)})' \in R^{K \times K}$ **and** $\beta^{(s)} \in (R^{+*})^K$



pdf of $W \sim N(-1.88, 3.78)$ **pdf of** $Y = max(W, 0)^{1.81}$

- **Assumptions on the covariance matrices**

  - **<u>HMMCI model</u> ($Q^2 - 1 + 3QK$ parameters)**

$$\Sigma^{(s)}(i,j) = \sigma_i^{(s)}\sigma_j^{(s)}\delta_{i,j} \text{ with } \sigma_j^{(s)} > 0$$

  - **<u>HMMdist model</u> ($Q^2 - 1 + K(3Q + 1)$ parameters)**

$$\Sigma^{(s)}(i,j) = \sigma_i^{(s)}\sigma_j^{(s)} exp(-\lambda^{(s)} d(z_i, z_j)) \text{ with } \sigma_j^{(s)} > 0 \text{ and } \lambda^{(s)} > 0$$

  - **<u>HMMloc model</u> ($Q^2 - 1 + 4QK$ parameters)**

$$\Sigma^{(s)}(i,j) = \sigma_i^{(s)}\sigma_j^{(s)} exp(-\lambda_i^{(s)}\lambda_j^{(s)} d(z_i, z_j)) \text{ with } \sigma_j^{(s)} > 0 \text{ and } \lambda_i^{(s)} > 0$$

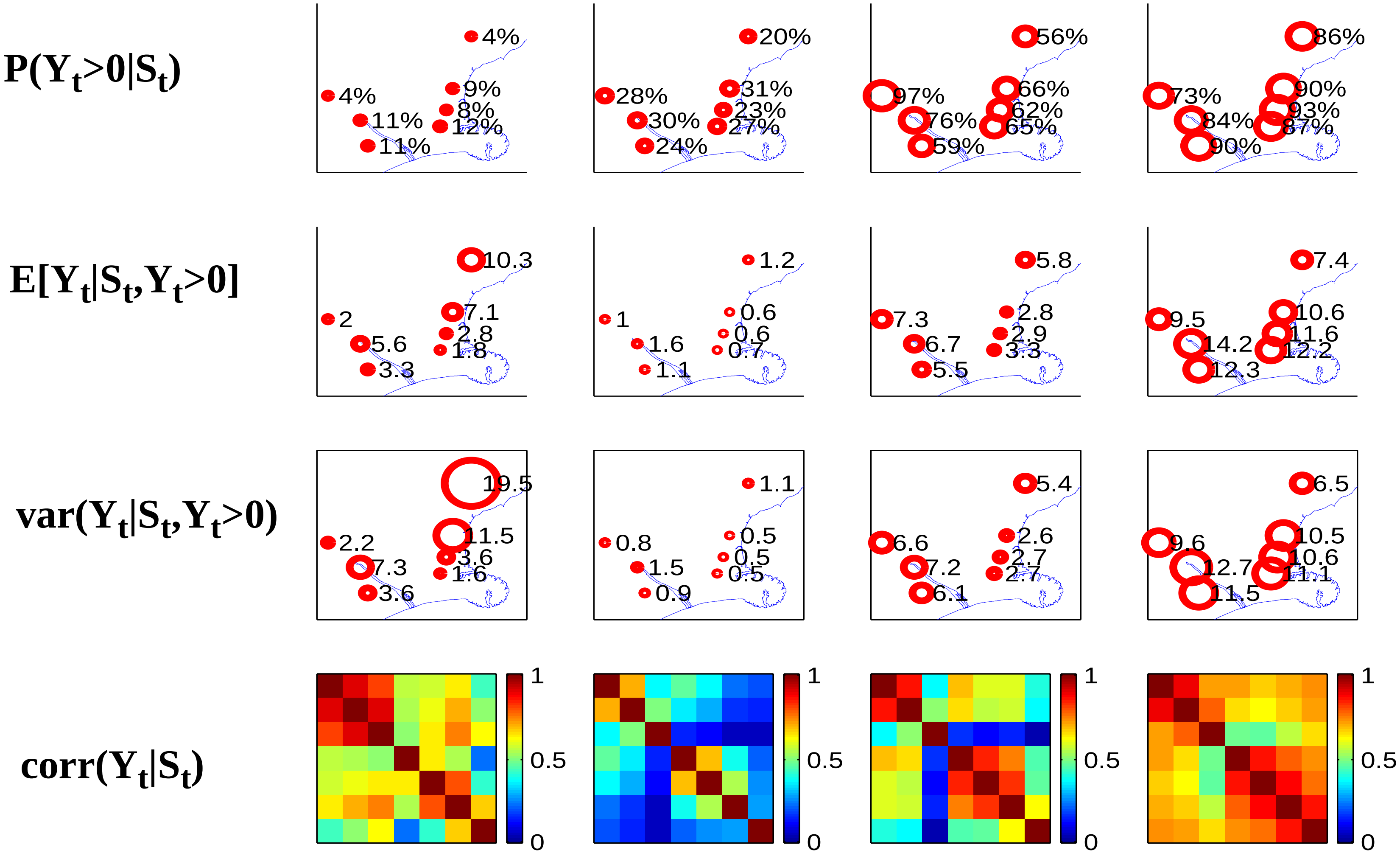
- **Model selection with AIC**

Q	**1**	**2**	**3**	**4**	**5**
Basic HMM	17404	14317	13436	13213	13144
HMMCI	17403	14445	13639	13398	13289
HMMdist	13092	12770	12697	12616	12623
HMMloc	12995	12741	12600	**12506**	12509
HMMfull	12904	12643	12640	12674	12611

Model validation

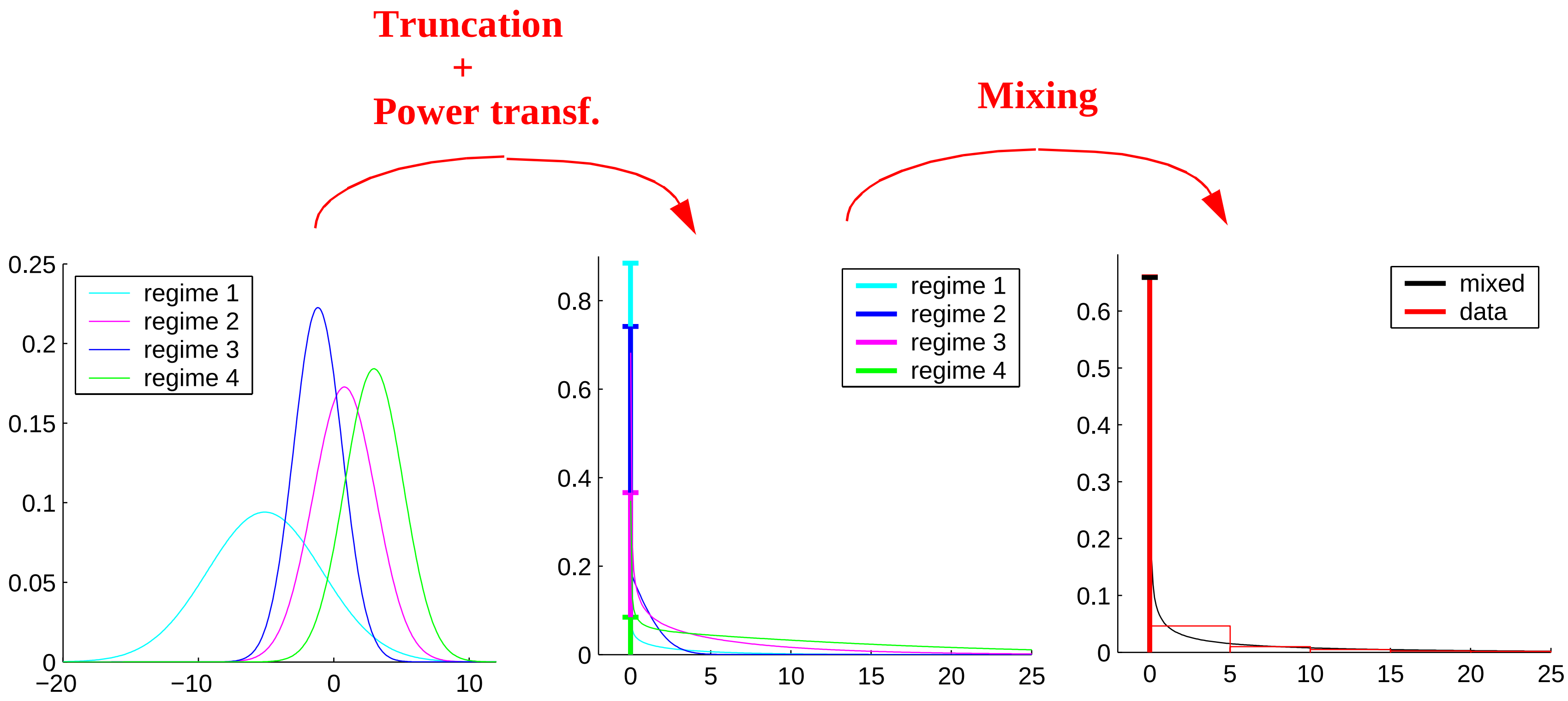
- **Meteorological interpretability**

Regime 1 **Regime 2** **Regime 3** **Regime 4**



Model validation

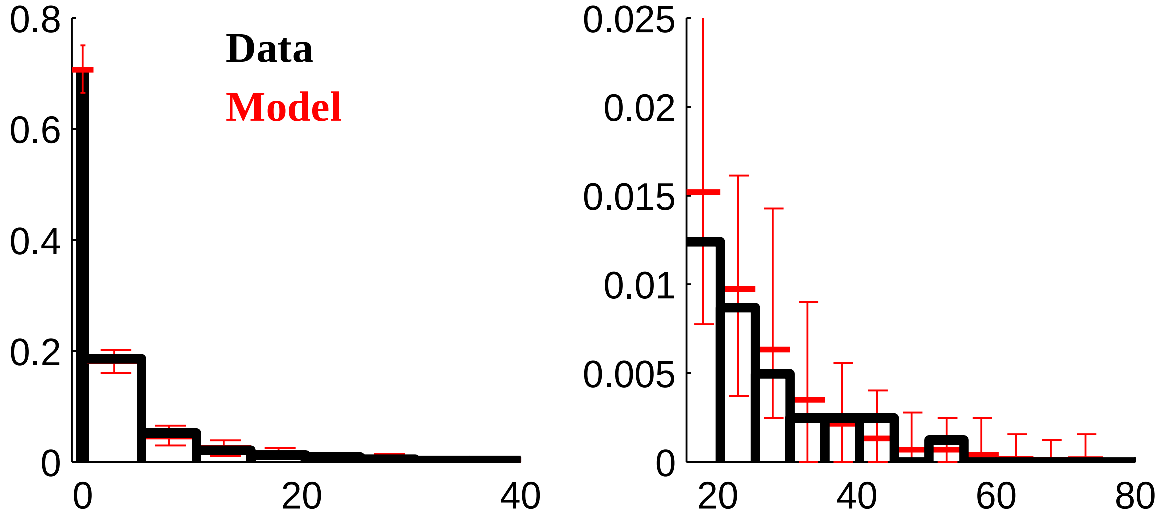
- **Meteorological interpretability: OK**
- **Realism of artificial sequences simulated with the model**
    - **Marginal distribution (Location 1)**



$$P(W_t(1)|S_t = s) \qquad P(Y_t(1)|S_t = s) \qquad \sum P(Y_t(1)|S_t = s)P(S_t = s)$$

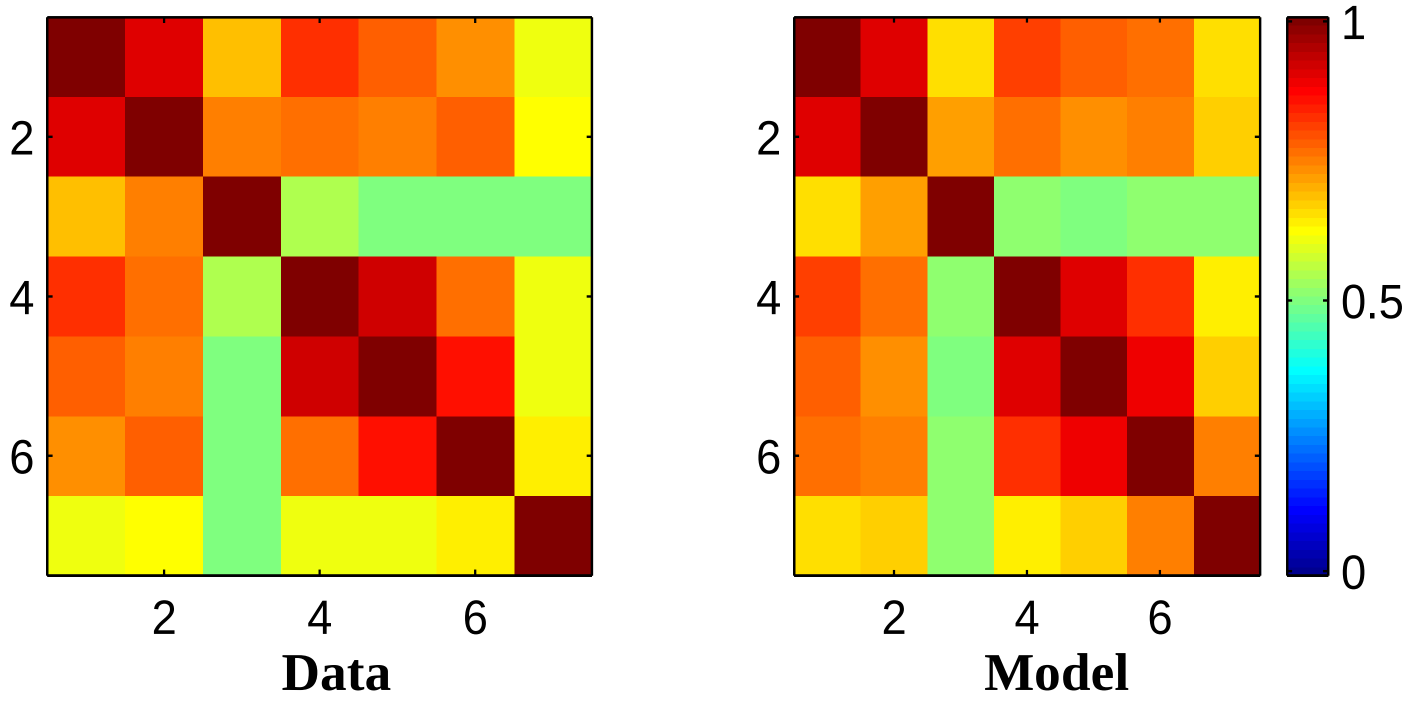
Model validation

- **Meteorological interpretability: OK**
- **Realism of artificial sequences simulated with the model**
  - **Marginal distribution (Location 1)**



Model validation

- **Meteorological interpretability: OK**
- **Realism of artificial sequences simulated with the model**
  - **Marginal distribution: OK**
  - **Dynamics at the different locations: ~OK**
  - **Spatial structure: OK**



Parameter estimation

- **Model structure**

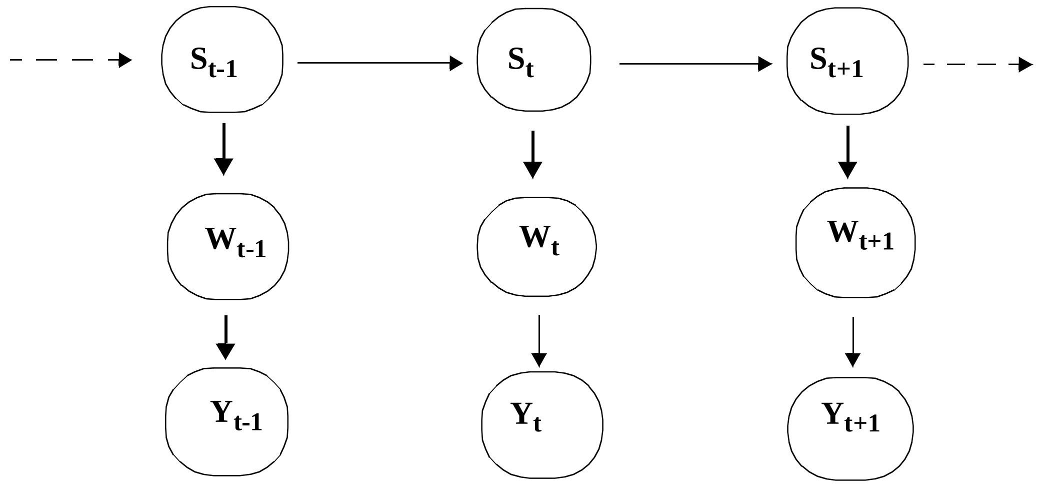
Weather type:
Not observed, finite values

Conditionally Gaussian vector:

Not observed, continuous values

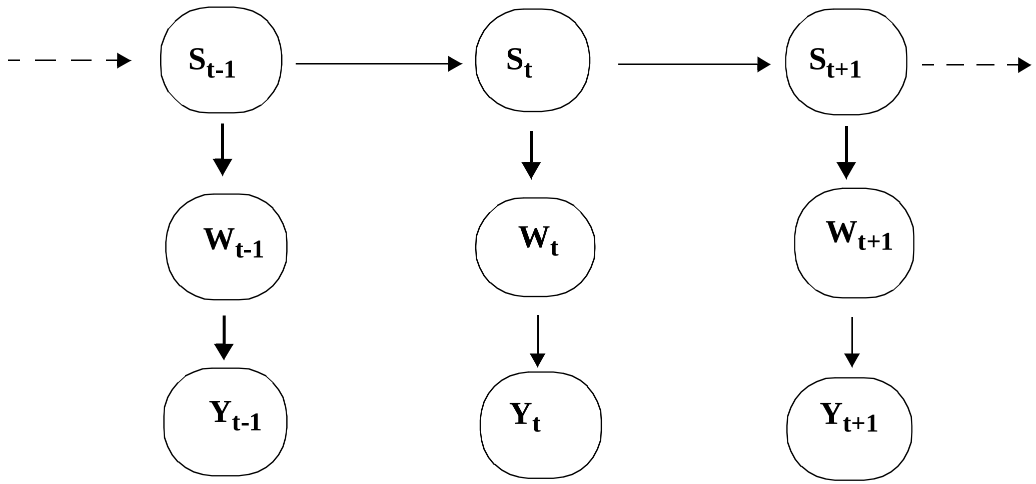
Precipitation:

Observed, mixed continuous-discrete



- **Monte-Carlo EM algorithm**

- **Need to compute the following smoothing probabilities for the M-step**
  - $u_t(s) = p(S_t = s \mid y_1^T, \theta_n)$
  - $v_t(s, s') = p(S_{t-1} = s, S_t = s' \mid y_1^T, \theta_n)$
  - $E[W_t \mid y_1^T, S_t = s, \theta_n]$
  - $E[W_t W_t' \mid S_t = s, y_1^T, \theta_n]$
- **Several algorithms can be used**
  - **Generic algorithms : Gibbs sampler, particle filter,...**
  - **More efficient algorithms if we take advantage of the specific structure of the model**

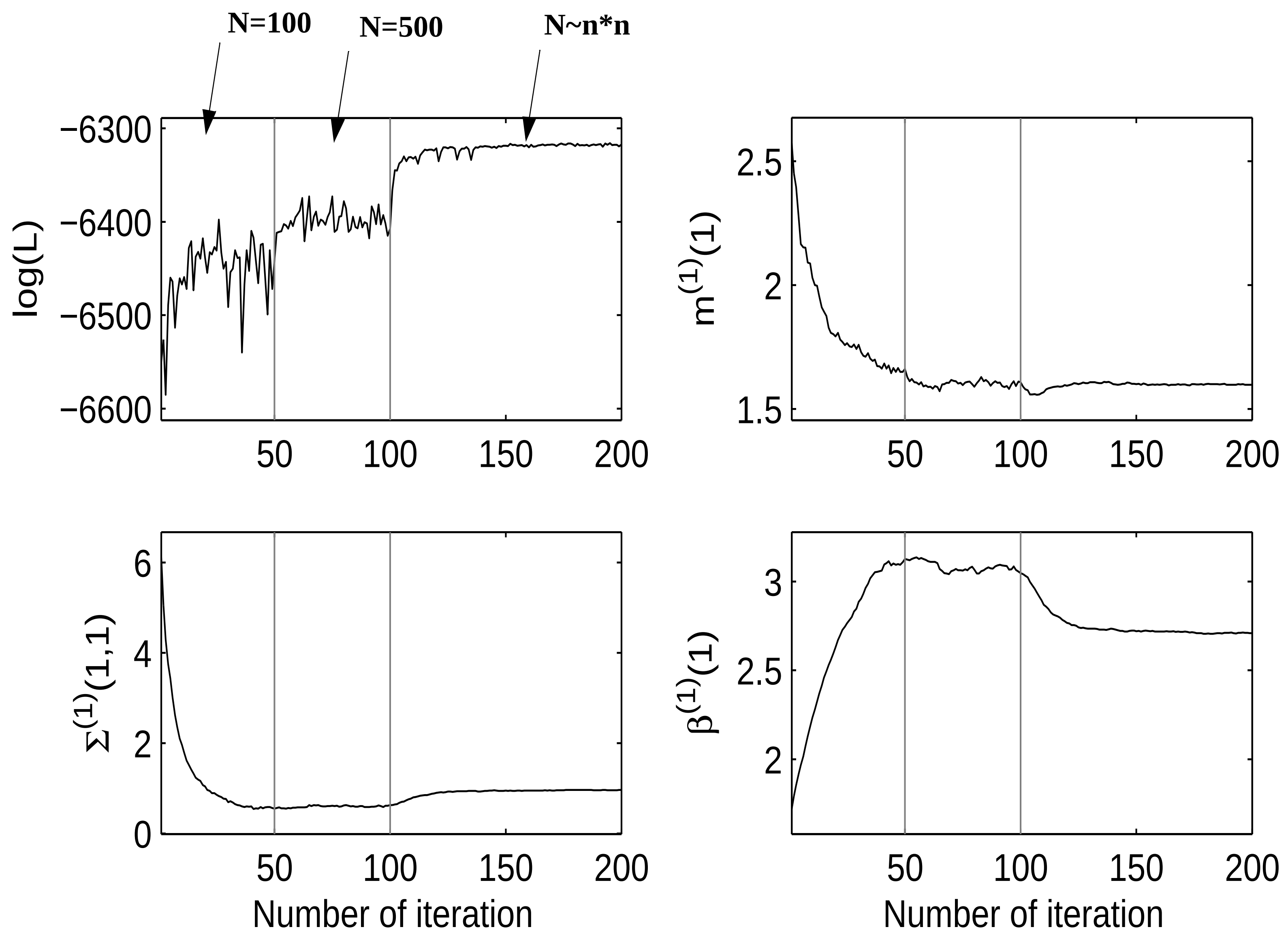


- **Need to compute the following smoothing probabilities for the M-step**
  - $u_t(s) = p(S_t = s|y_1^T, \theta_n)$
  - $v_t(s, s') = p(S_{t-1} = s, S_t = s'|y_1^T, \theta_n)$
  - $E[W_t|y_1^T, \theta_n]$
  - $E[W_t W_t'|y_1^T, \theta_n]$
- **Several algorithms can be used**
  - **Generic algorithms : Gibbs sampler, particle filter,...**
  - **More efficient algorithms if we take advantage of the specific structure of the model**
    - **Forward-backward algorithm for the discrete component**
    - **Monte-Carlo integration for the gaussian component (ϕ pdf of $N(m, \Sigma)$)**

$$\int\limits_{]-\infty, 0]^k} \int \phi(w_1, \ldots, w_k, w_{k+1}, \ldots, w_K) dw_1 \ldots dw_k$$

$$\int\limits_{]-\infty, 0]^k} \int w_i \phi(w_1, \ldots, w_k, w_{k+1}, \ldots, w_K) dw_1 \ldots dw_k$$

$$\int\limits_{]-\infty, 0]^k} \int w_i w_j \phi(w_1, \ldots, w_k, w_{k+1}, \ldots, w_K) dw_1 \ldots dw_k$$


N=100
N=500
N~n*n
log(L)
−6300
−6400
−6500
−6600
50
100
150
200
m(1)(1)
2.5
2
1.5
Σ(1)(1,1)
6
4
2
0
β(1)(1)
3
2.5
2
Number of iteration
Number of iteration

Conclusion

- **Better description of the spatial structure of the data**
- **Some perspectives**
  - **Add an autoregressive part to better describe the local dynamics**
  - **Incorporate seasonal and inter-annual components in the model**
- **Some virtues of HMM**
  - **Distributional versatility**
  - **Ability to model diverse time scales**
  - **Open structure which allows for more physical models**