UNIVERSITAT POLITÈCNICA DE VALÈNCIA

iiama Instituto de Ingeniería del Agua y Medio Ambiente

IMPROVING THE BAYESIAN JOINT INFERENCE THROUGH THE INCLUSION OF HYDROLOGICAL STATE VARIABLES IN THE RESIDUALS DEPENDENCE MODEL

M.R. Hernández and F. Francés

Research Group on Hydrologic and Environmental Modeling
Research Institute of Water and Environmental Engineering
***Universitat Politècnica de València**, Spain*

26th IUGG GENERAL ASSEMBLY 2015
INTERNATIONAL UNION OF GEODESY AND GEOPHYSICS
UNION GÉODÉSIQUE ET GÉOPHYSIQUE INTERNATIONALE

PRAGUE, CZECH REPUBLIC
PRAGUE CONGRESS CENTRE
JUNE 22–JULY 2, 2015
Earth and Environmental Sciences for Future Generations

Introduction (1/3)

- **Problem:** Hydrological models provide predictions, which are not lacking of uncertainty
  - In general, model state variables(e.g. streamflow "q_s") do not match observations of the predictand "q" $q \neq q_s$
- Considering
  - "q_s" as a Random Variable
  - The existence of **the joint pdf** $p\left(q, q_s(\theta_{h,e}; \tilde{X}_n)\right)$

  - We can define the **Predictand pdf conditioned on q_s (Predictand cpdf)** $p\left(q \middle| q_s\right)$
- So far, equations are independent of the kind of error model (additive/multiplicative)
  - If we consider an **additive error**, **Predictand cpdf = Error cpdf**

$$q = q_s + e \quad \rightarrow \quad p\left(q \middle| q_s\right) = p\left(e \middle| q_s\right)$$

Introduction (2/3)

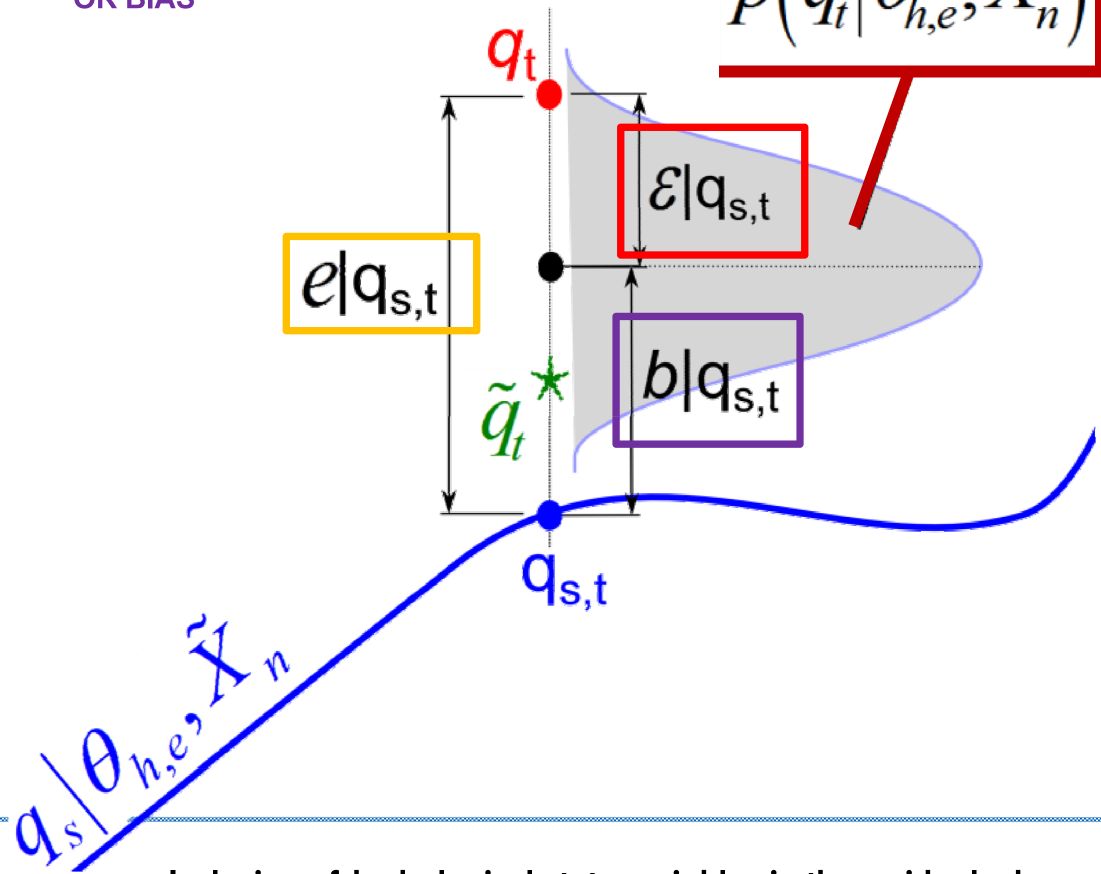
- **Modeling the Error term** $q = q_s + e$
  - We can model the **two components of Error** at time step "t"

$$q_t = q_{s,t} + b|q_{s,t} + \varepsilon|q_{s,t} \rightarrow p\left(q_t|q_{s,t}\right) = p\left(e|q_{s,t}\right) = p\left(\varepsilon|q_{s,t}\right)$$

DETERMINISTIC OR BIAS RANDOM ERROR

$$p\left(q_t|\theta_{h,e},\tilde{X}_n\right)$$

THE PREDICTIVE CPDF



Introduction (3/3)

- **Classical approach for modeling the Error term**
  - Considers additive errors serially uncorrelated (White Noise)
  - With Gaussian distribution
  - Constant conditional variances (homoscedastic errors)
  - It does not account for Bias

UNBIASED I.I.D. ERROR

➢ **Equivalent to Std. Least Squares calibration (SLS)**

- **Errors in Hydrology** do not satisfy the **SLS** hypotheses
  - **Causes** are mainly the **Input errors** and an unsuitable **H. Model structure**
  - **Consequences**
    - Biased or "corrupted" **parameter values**
    - An incorrect estimation of the **Predictive uncertainty**

Aims of this research

- **Phase I**
  - Inferring a **Specific Error Model** that best fits Hydrological Model Errors
    - **Inference must be a JOINT INFERENCE to avoid Biased parameters in both models**
  - Compare Performance of **SLS vs Specific Error Model**

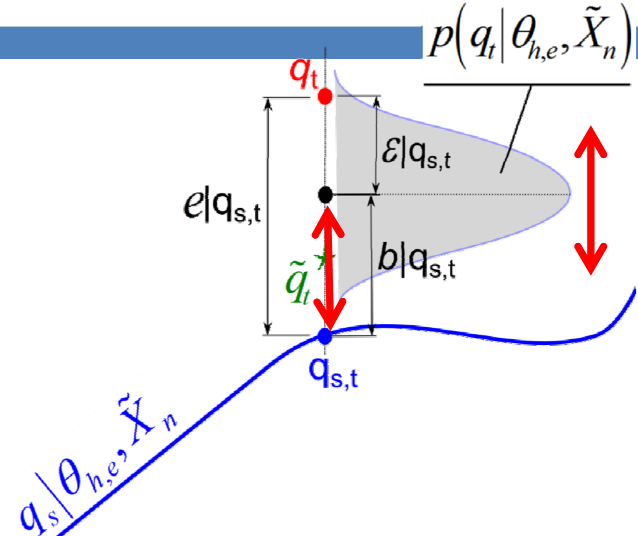
- **Phase II** (Not concluded)
  - We try an upgrade of the error model through a **Bias model improvement**, in order to achieve a better performance than in phase I

Phase I: Error Model description

Time-varying Error variance & Bias

Variance $\sigma_{\varepsilon|q_{s,t}} = \theta_1^e + \theta_2^e q_{s,t}$

Bias $\begin{cases} b_{e|q_{s,t}} = \theta_3^e + \theta_4^e q_{s,t} & q_{s,t} \le \theta_5^e \\ b_{e|q_{s,t}} = \theta_6^e + \theta_7^e q_{s,t} & q_{s,t} > \theta_5^e \end{cases}$



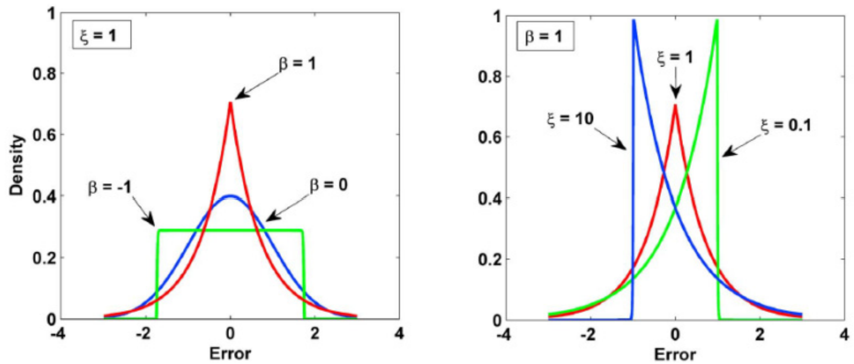
Modeling the Errors dependence
through an **AR(p)** model

Unbiased Studentized Error $\eta_t = \dfrac{e|q_{s,t} - b|q_{s,t}}{\sigma_{\varepsilon|q_{s,t}}} = \dfrac{\varepsilon|q_{s,t}}{\sigma_{\varepsilon|q_{s,t}}}$

$$\eta_t = \sum_{i=1}^{p} \phi_i \eta_{t-i} + \boxed{z_t} \equiv \phi_p(B)\eta_t = z_t$$

Innovations (White Noise)

Modeling innovations Z_t through the flexible Skew Exponential Distribution (**SEP**)



[Schoups and Vrugt (2010)]

Symmetric and Gaussian as particular cases

$$q_t = q_{s,t} + b|q_{s,t} + \sigma_{\varepsilon|q_{s,t}}\left[\phi_p^{-1}(B)[Z_t]\right] \rightarrow p\left(q_t|q_{s,t}\right) = p\left(e|q_{s,t}\right) = p\left(\varepsilon|q_{s,t}\right)$$

Phase I: Total Laws

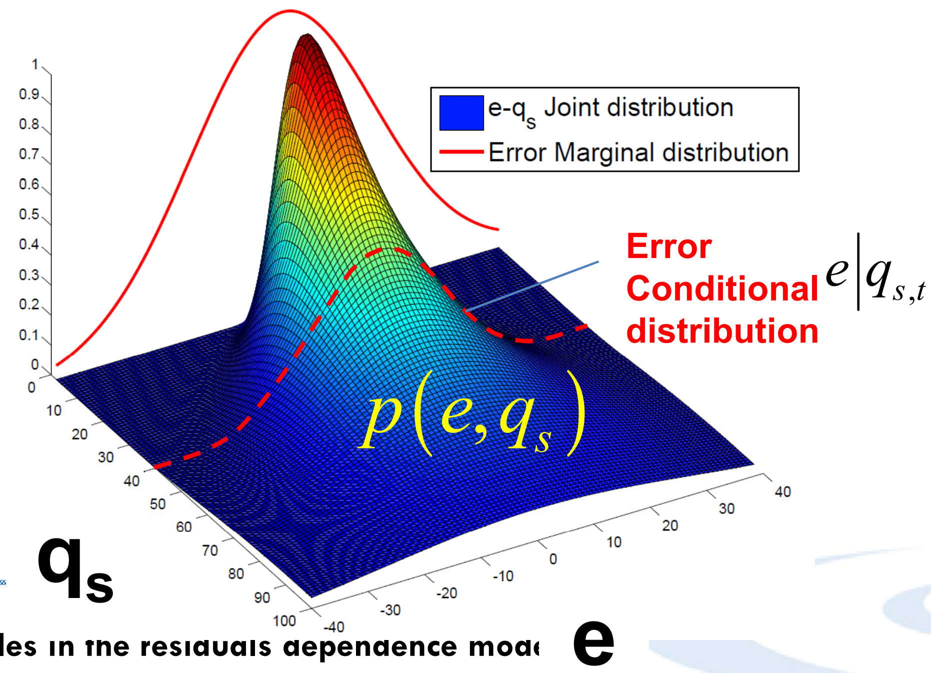
- **In defined Error model, some parameters of variance & Bias functions are not free !**
  - **Marginal and Conditional** Error distributions belong to the **same joint** distribution $p(e, q_s)$
    - Linked by Total Variance Law (**TVL**) and Total Expectation Law (**TEL**)
    - For the correct implementation of the **JOINT INFERENCE with a Time-Varying Error Model ➔ TOTAL LAWS** must be enforced **!**

TVL

$$V(\mathbf{e}) = E\left[V\left(e|q_{s,t}\right)\right] + V\left[E\left(e|q_{s,t}\right)\right]$$

TEL

$$E(\mathbf{e}) = E\left[E\left(e|q_{s,t}\right)\right] = E\left[b|q_{s,t}\right]$$



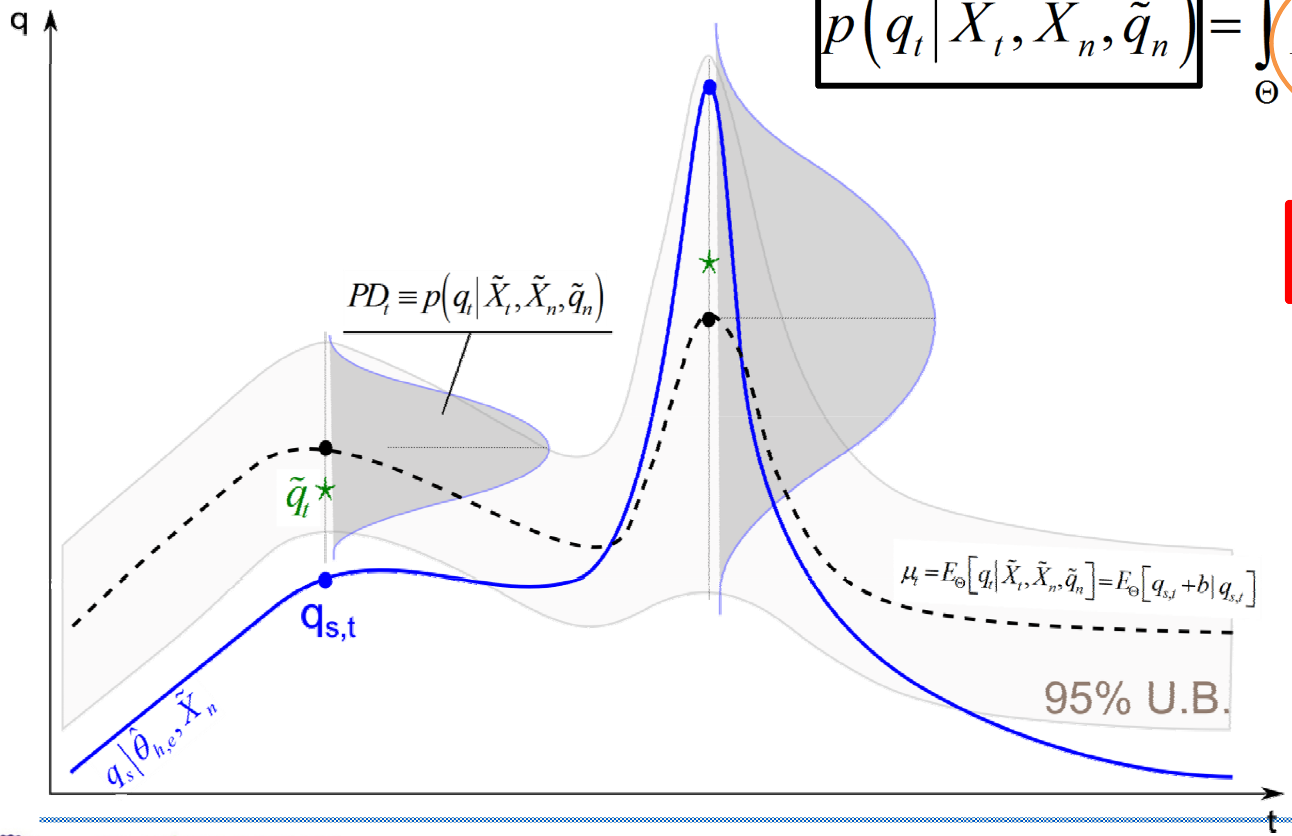
Phase I: Predictive uncertainty

Given the previously obtained, **Predictive pdf conditioned** on the simulated streamflow…

$$p\left(q_t | q_{s,t}\right) = p\left(e | q_{s,t}\right) = p\left(\varepsilon | q_{s,t}\right)$$

…we can get the **Predictive pdf** by its Marginalization on the parameters…

$$p\left(q_t | \tilde{X}_t, \tilde{X}_n, \tilde{q}_n\right) = \int_\Theta p\left(q_t | \theta_{h,e}, \tilde{X}_n\right) g\left(\theta_{h,e} | \tilde{q}_n, \tilde{X}_n\right) d\theta$$

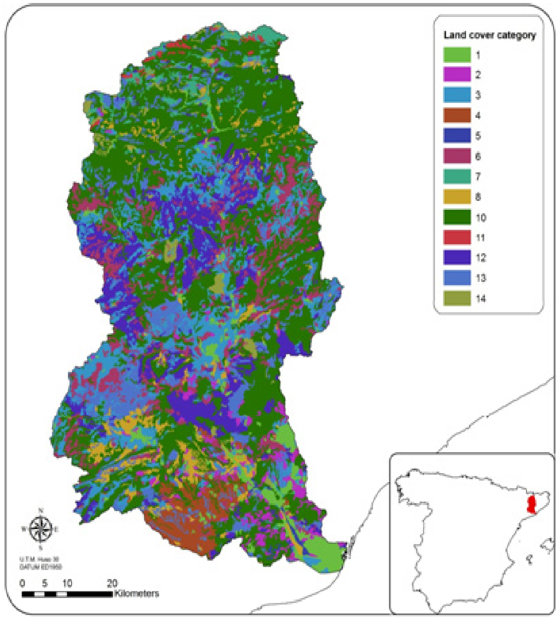


Posterior of parameters

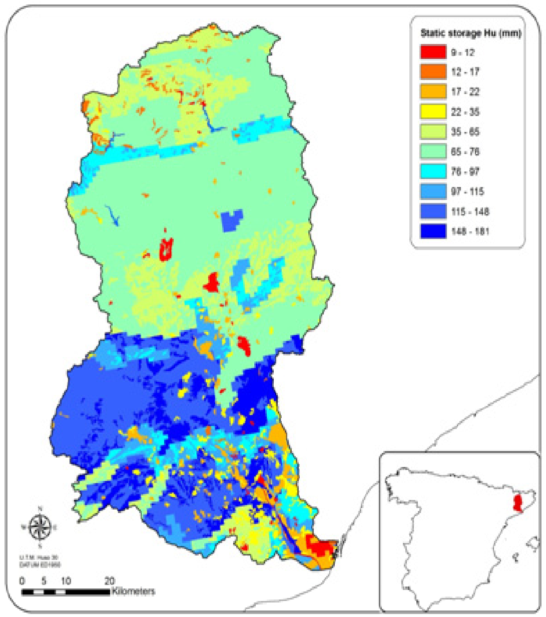
- Bayesian **Joint** inference
- MCMC sampling
  - ***DREAM-ZS*** algorithm
    [*Ter Braak and Vrugt (2008)*]

Phase I: Case Study

- **Distributed Hydrological Model (on a Spanish humid catch.)**
  - **TETIS** http://lluvia.dihma.upv.es/EN/software/software.html
  - ➢ Effective Parameter Structure divided in two parts:
    - An estimated Value in each cell setting-up the **Parameter Maps**
    - **Regularization Function**: **Global calibrated correction factor F_i** applied to each parameter map



x F1



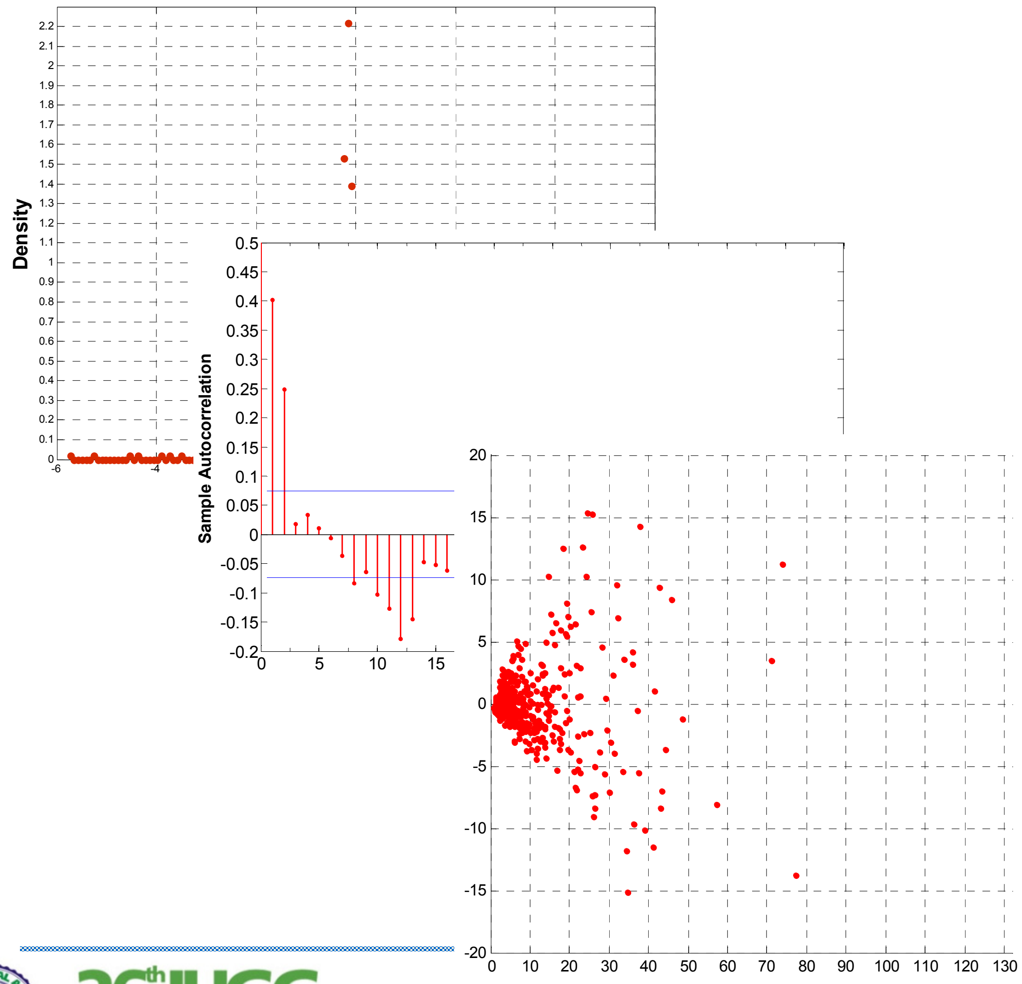
x F2 …….

Phase I: SLS vs EM2 comparison (1/5)

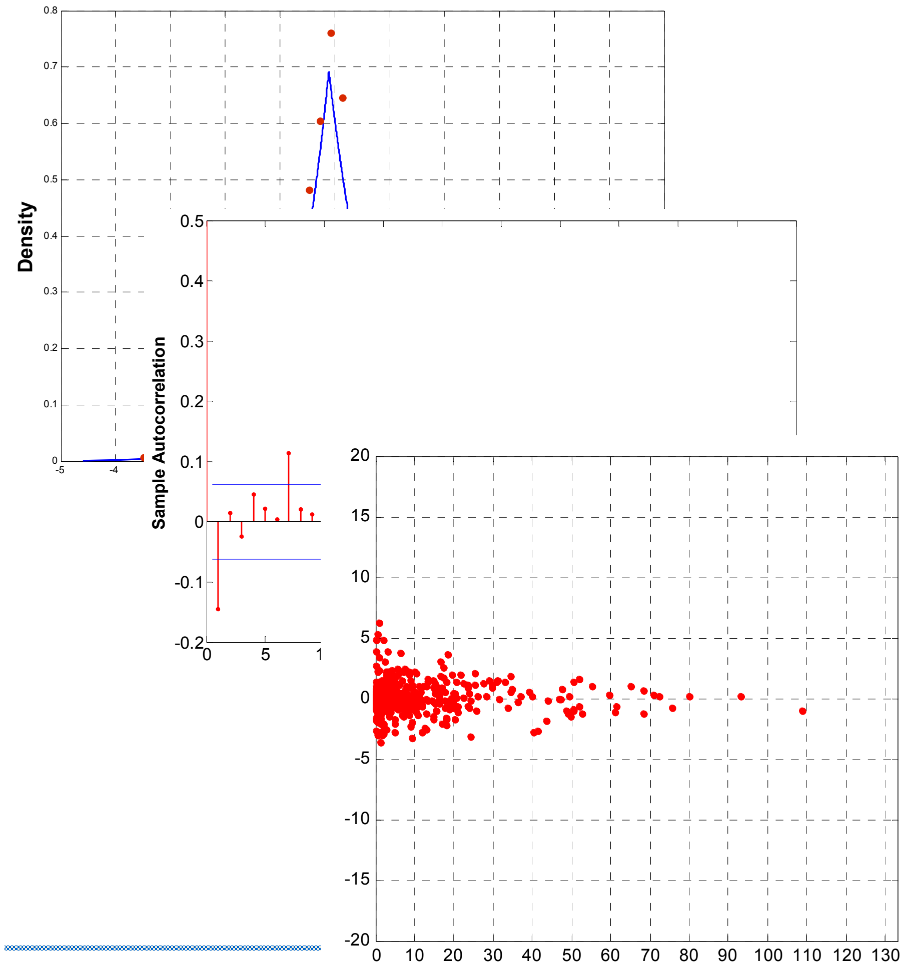
- **Fulfillment of the Error Model Hypothesis**

Normality ***Independence*** ***Homoscedasticity***

SLS



EM2



Inclusion of hydrological state variables in the residuals dependence model

Phase I: SLS vs EM2 comparison (2/5)

- **Simulation performance**
  - **In our Case Study** both show a **similar performance** of prediction in Validation based on NSE, RMSE, and VE% indexes

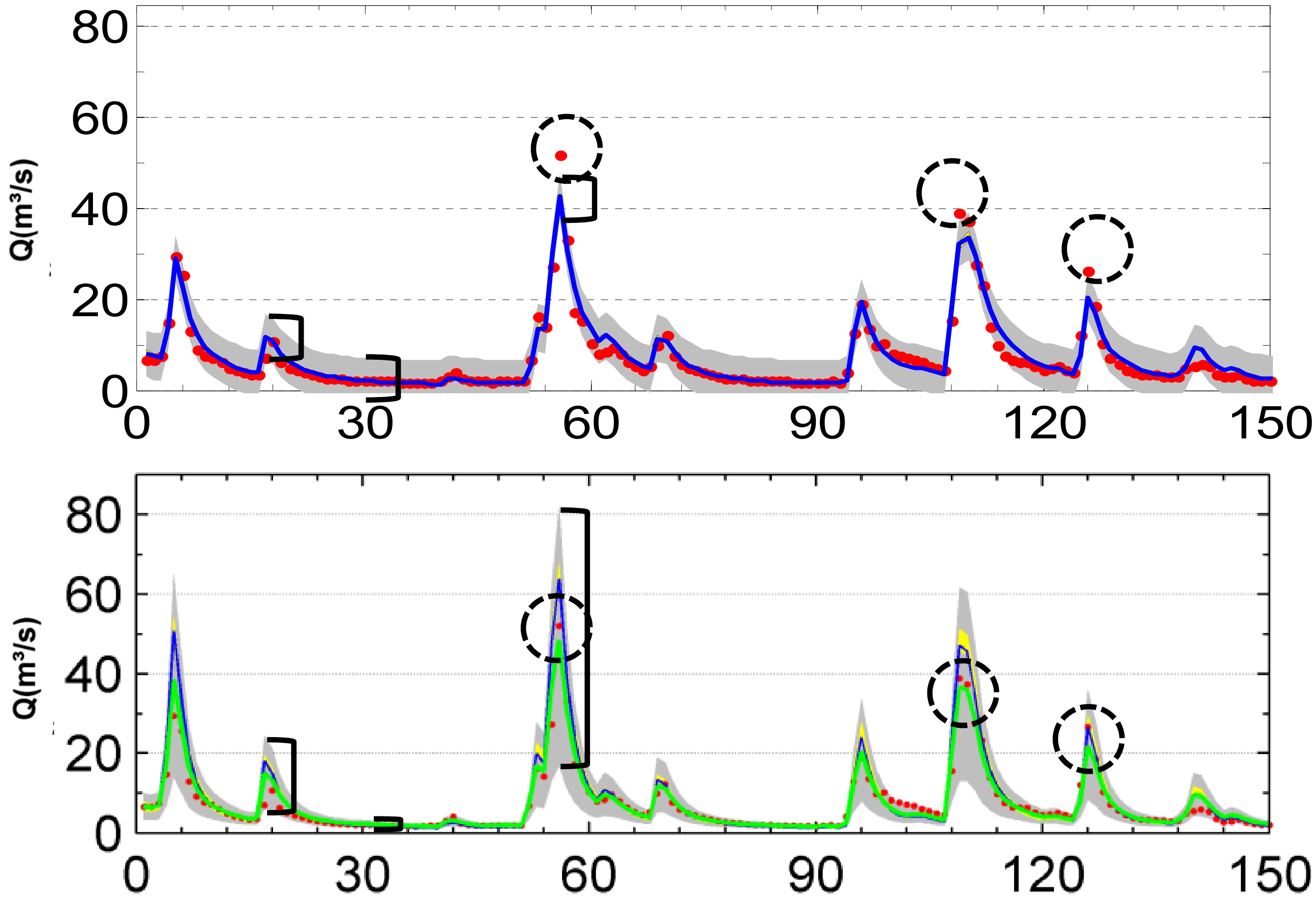
		SLS			EM2		
		CALIB	VALID	% CHANGE	CALIB	VALID	% CHANGE
HYDRO MODEL	NSE	0.93	0.86	7%	0.74	0.72	3%
	RMSE	2.62	3.48	33%	5.00	4.99	0%
	ErrVol (%)	2.40	-4.5	88%	9.90	2.70	73%
MEAN PREDICTION	NSE				0.91	0.85	7%
	RMSE				2.92	3.60	23%
	ErrVol (%)				0.01	-3.70	

Phase I: SLS vs EM2 comparison (3/5)

- **Assessment of the Predictive Uncertainty**
  - **95% Uncertainty Band**

SLS

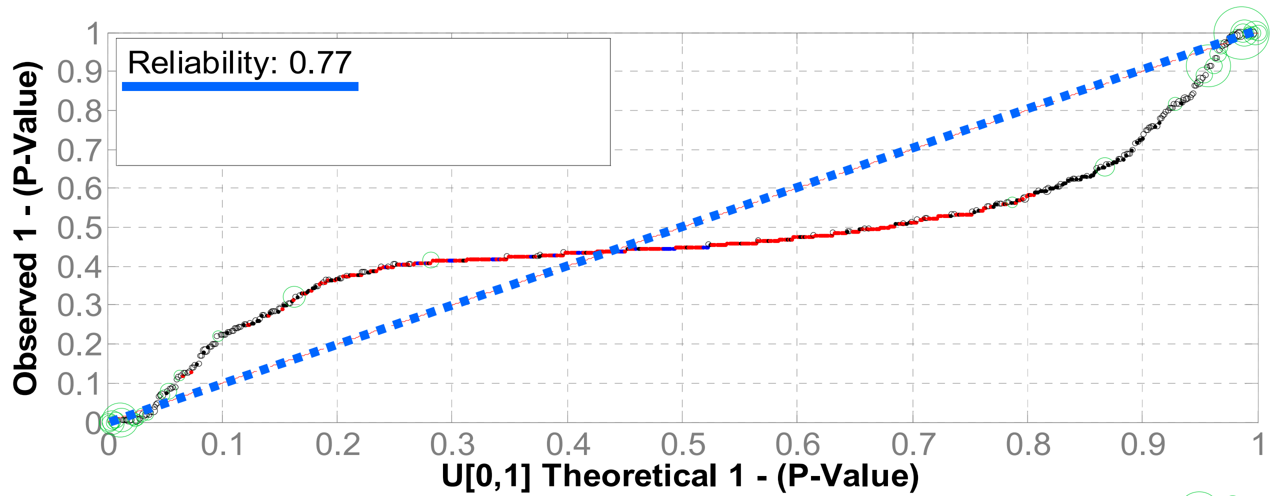
EM2



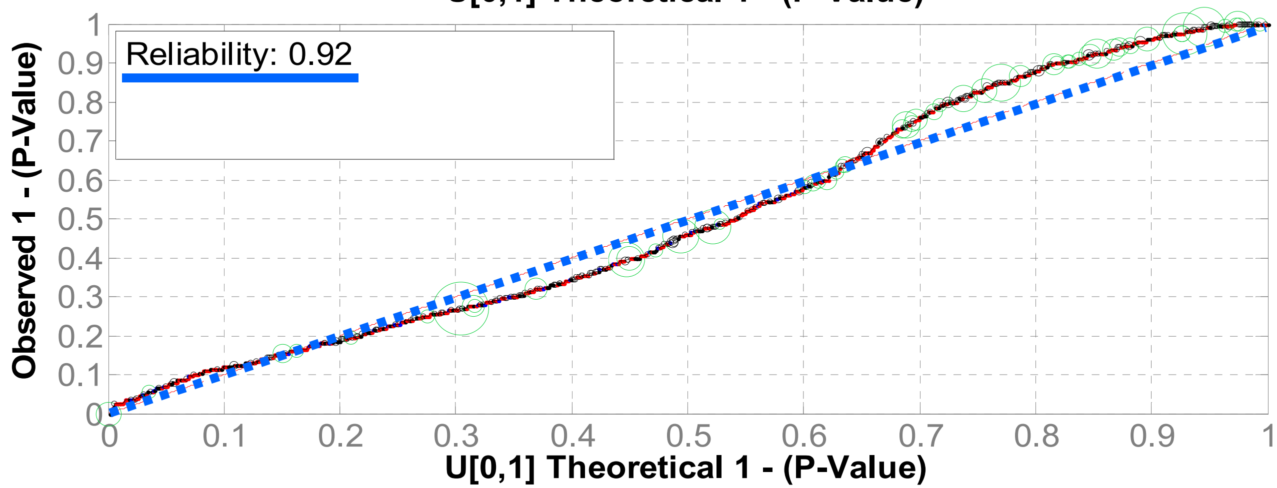
Phase I: SLS vs EM2 comparison (4/5)

- Assessment of the Predictive Uncertainty
  - Full Predictive distribution Reliability (PP-PLOTS)

SLS



EM2



Phase I: SLS vs EM2 comparison (5/5)

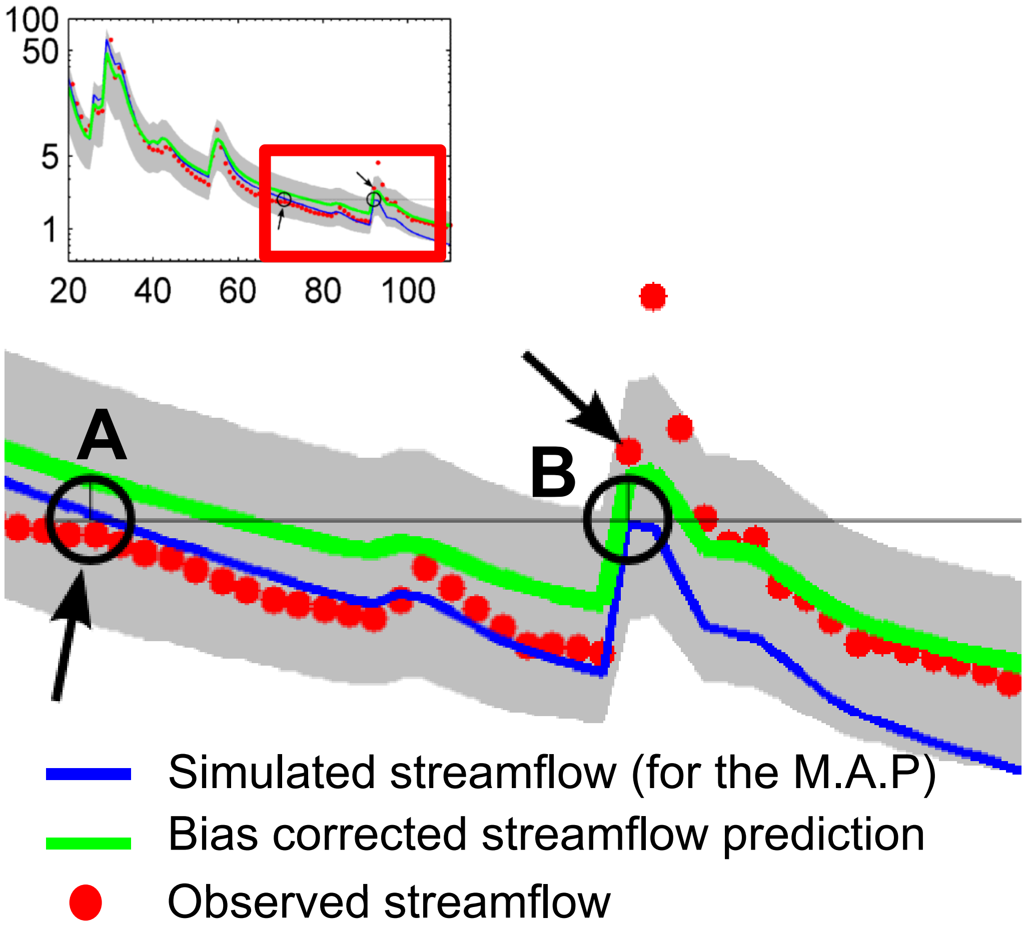
- **Parameters value coherence**
  - **EM2 shows less Biased parameters than SLS**
    - EM2 exhibits less deterioration of the H. Model performance between calibration and validation (**Divergence Phenomenon**)

		SLS			EM2		
		CALIB	VALID	% CHANGE	CALIB	VALID	% CHANGE
HYDRO MODEL	NSE	0.93	0.86	7%	0.74	0.72	3%
	RMSE	2.62	3.48	33%	5.00	4.99	0%
	ErrVol (%)	2.40	-4.5	88%	9.90	2.70	73%

- For some parameters SLS yields inferred values without **physical meaning**

Motive of Phase II

- **Phase I:** Conditional Bias ←==→ q_s **a biunivocal relation**
  - it doesn't show a good performance in validation



- Point **A**: at streamflow recession scenario
  Point **B**: at streamflow peak scenario
  - Both points show same "q_s" (**Blue L.**) and therefore same Bias (**Green L.**)
  - Observations $\tilde{q}_t$ (**Red points**)
    - At A, below "q_s"
    - At B, above "q_s"

> Different hydrological scenarios with different active processes, can yield **the same "q_s"** value

> But H. Model can produce **different Bias** when it models those different hydrological scenarios

Phase II: Bias Model description (1/2)

- In classical linear regression is common the analysis of residuals searching for a **misspecification of the fitted model** $e = f\left(regressors\right)$**?**
- A fitted H. model with **structural problems** could also exhibit **residuals correlated with "exogenous" variables (regressors)** of the modeled processes …
  - Variables to include ? (all Standardized $\mathbf{X}_k$ k=1,…,n):
    - **State Variables: Runoff** (t), **Interflow**(t) and **Base Flow** (t, t-1, t-2, …)
    - Or even, **Forcing: Precipitation** (t, t-1, t-2, …)
  - Residuals dependence model: **ARX(p,q)**

$$\eta_t = \sum_{i=1}^{p} \phi_i \eta_{t-i} + \sum_{k=1}^{n} \sum_{j=0}^{q} W_{kj} X_{k,t-j} + Z_{arx,t} \equiv \phi_p\left(B\right)\boldsymbol{\eta} = \sum_{k=1}^{n} W_{kq}\left(B\right)\mathbf{X}_k + \mathbf{Z}_{arx}$$

Studentized Error — Autoregressive Filter — "Exogenous" Filter — Innovations (White Noise)

Phase II: Bias Model description (2/2)

- **Relation between ARX(p , q) model and the Bias model**

$$\phi_p(B)\,\boldsymbol{\eta} = \sum_{k=1}^{n} W_{kq}(B)\,\mathbf{X}_k + \mathbf{Z}_{arx}$$

$$\boldsymbol{\eta} = \phi_p^{-1}(B)\left[\sum_{k=1}^{n} W_{kq}(B)\,\mathbf{X}_k + \mathbf{Z}_{arx}\right] = \underbrace{\phi_p^{-1}(B)\left[\sum_{k=1}^{n} W_{kq}(B)\,\mathbf{X}_k\right]}_{\boldsymbol{\eta}_I} + \underbrace{\phi_p^{-1}(B)\left[\mathbf{Z}_{arx}\right]}_{\boldsymbol{\eta}_{II}}$$

$\boldsymbol{\eta}_I$ Explained part of Studentized Error

$\boldsymbol{\eta}_{II}$ Random part of Studentized Error

$$\boldsymbol{\eta} = \frac{\mathbf{e}}{\boldsymbol{\sigma}_{e|q}}$$

Studentized Error

$$\boldsymbol{\eta} = \frac{\mathbf{e}}{\boldsymbol{\sigma}_{e|q}} = \boldsymbol{\eta}_I + \boldsymbol{\eta}_{II}$$

$$\frac{\mathbf{e} - \boldsymbol{\eta}_I \boldsymbol{\sigma}_{e|q}}{\boldsymbol{\sigma}_{e|q}} = \boldsymbol{\eta}_{II}$$

UNBIASED Studentized Error

$$E[\boldsymbol{\eta}_{II}] = 0 \qquad V[\boldsymbol{\eta}_{II}] = 1$$

- $\boldsymbol{\eta}_I \boldsymbol{\sigma}_{e|q}$: Bias model is a function of "n" $\mathbf{X}_k$ **Std. State Variables**
  - $\mathbf{W}_{kq}$ { are regression coefficients to be **inferred JOINTLY**; offer us keys about the **sources of the Bias** (research!)

iiama

Phase II: *"Man is the only animal to stumble twice on the same stone"* *Spanish proverb*

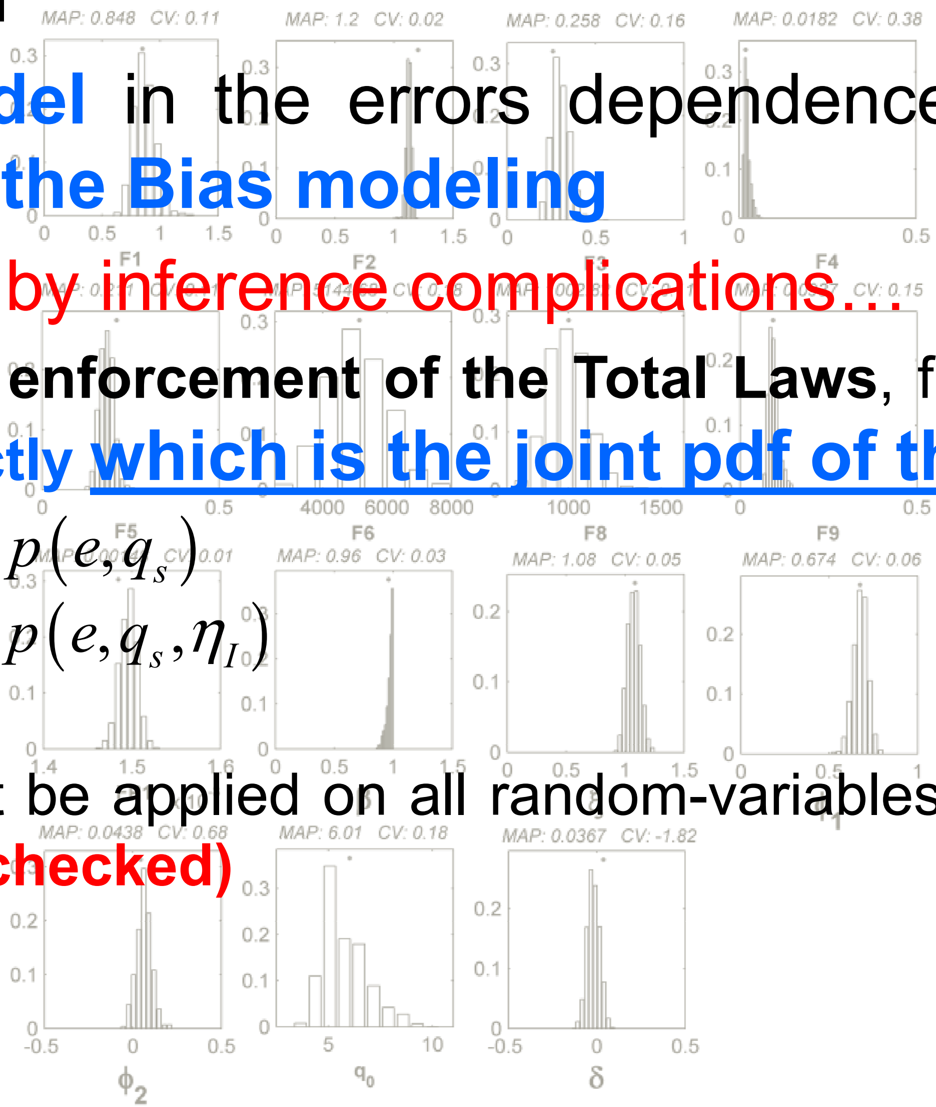
- **Inference complications in Phase II**
  - Markov chains do not converge (> 1.000.000 Iter.)
  - Inferred lag(1) autocorrelation parameter too high (~ 0.99)
  - Similar problems reported in papers where **TL** were not applied (e.g. [*Schoups and Vrugt (2010)*] [Evin et al. *(2013, 2014)*])
- **Suspected origin of the problem:**
  - **In Phase I** we had a **bivariate pdf** $p(e, q_s)$ on which **TL** were applied, and inference was **successful**
  - **In Phase II** we have a **trivariate pdf** $p(e, q_s, \eta_I)$ on which **TL** is applied but only on $p(e, q_s)$ and inference has **failed…**
    - ***Total Laws*** must be enforced also,
      - on $p(e, \eta_I)$ and on $p(q_s, \eta_I)$ **?**

Conclusions

- For the **Phase I**
  - In our **Case Study**, Error models performance shows that
    - SLS doesn't fulfill **the initial hypotheses** while the EM2 fulfillment is good
    - **Simulations in validation** show that EM2 and SLS have a **similar performance** (**NSE** & **Vol.Error**)
    - EM2 yields **less biased hydrological parameters**, which is interesting for Regionalization methods
    - EM2 shows **better** assessment of the **Predictive Uncertainty**
  - **Bias model** needs to consider **more explicative variables** than the simulated streamflow → Motive of Phase II
  - Time-Varying Error Models must enforce ***THE TOTAL LAWS*** (**TVL** and **TEL**)

Conclusions

- For the **Phase II**
  - An **ARX model** in the errors dependence model, it is proposed **for the Bias modeling**
  - Unconcluded by inference complications…
    - It seems that **enforcement of the Total Laws**, firstly requires to **define correctly which is the joint pdf of the Error**
      - Bivariate $p(e,q_s)$
      - Trivariate $p(e,q_s,\eta_I)$
      - …
    - If so, TL must be applied on all random-variables in that joint pdf **(issue to be checked)**



UNIVERSITAT POLITÈCNICA DE VALÈNCIA

iiama Instituto de Ingeniería del Agua y Medio Ambiente

Thank you

for your attention

Mario R. Hernández (maherlo@hma.upv.es)

Research Group on Hydrologic and Environmental Modeling

lluvia.dihma.upv.es

The study was funded by the Spanish Ministry of Economy and Competitiveness through the research projects SCARCE (CSD2009-00065) and ECOTETIS (CGL2011-28776-C02-01), and by the Universitat Politècnica de València through the Research and Development Grants Program (PAID).