Hydrological Ensemble Prediction EXperiment - 2007
STRESA - June 27-29, 2007

Reconciling hydrological physically based and data driven models in terms of predictive probability

E. Todini, G. Coccia
University of Bologna

C. Mazzetti
ProGeA Srl

SCOPE

This presentation aims at showing how Hydrological models of vastly different nature (from the Deterministic to the Data Driven ones), used in Flood Forecasting, can be reconciled in terms of

Predictive Probability

THE HYDROLOGICAL DEBATE

For many years, hydrologists have debated on the appropriateness of using Data Driven models as opposed to Conceptual or Physically Based models for Flood Forecasting and, in particular, for Real Time Flood Forecasting.

THE HYDROLOGICAL DEBATE

In 1983 Klemes advocated ways of combining both approaches in order to capitalize on the greater accuracy of the Data Driven models as well as on the larger forecasting stability provided by the Physically Based models.

Klemes, V. K., 1983. Conceptualization and scale in hydrology, *J. Hydrol.*, **65**,1-23

THE BAYESIAN PROCESSORS

In 1999 Krzysztofowicz with his "Bayesian theory of probabilistic forecasting via deterministic hydrologic models" opened the main streams of Bayesian post-processors, followed, more recently by Raftery et al. 2003, who introduced the Bayesian Model Averaging.

Krzysztofowicz, R., 1999. Bayesian theory of probabilistic forecasting via deterministic hydrologic model. *Water Resour. Res.*, **35**, 2739–2750.
Raftery, A. E., F. Balabdaoui, T. Gneiting, and M. Polakowski, 2003. Using Bayesian model averaging to calibrate forecast ensembles, *Tech. Rep. 440*, Dep. of Stat., Univ. of Wash., Seattle.

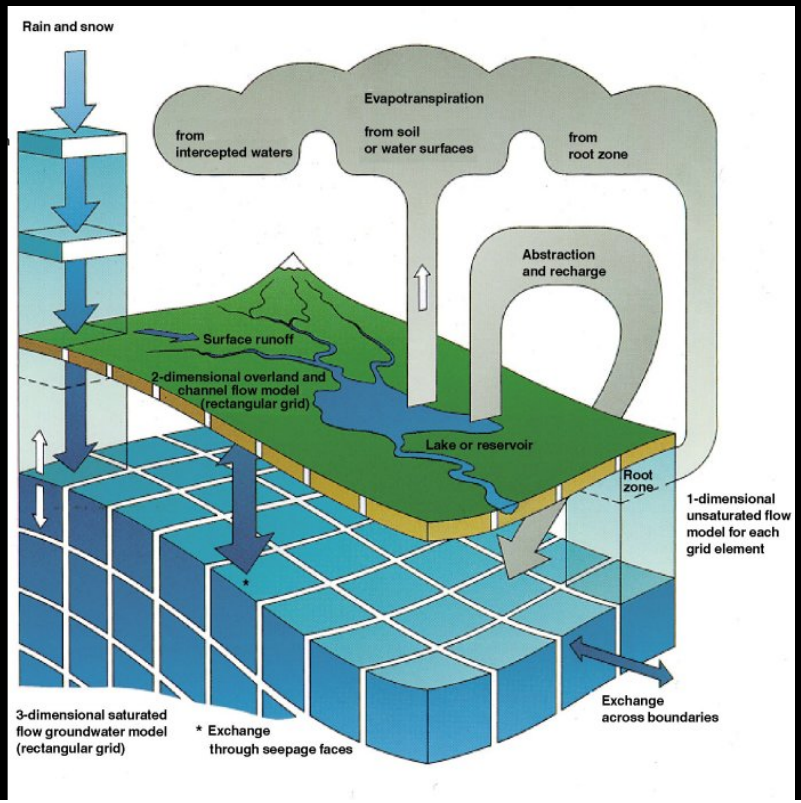
A NEW APPROACH

It is the ambition of this presentation to introduce a new post-processing approach which will allow to generalise the Krzysztofowicz results to multi-models of different nature, and, at the same time, to simplify the computational requirements of the Raftery's Bayesian Model Averaging processor.

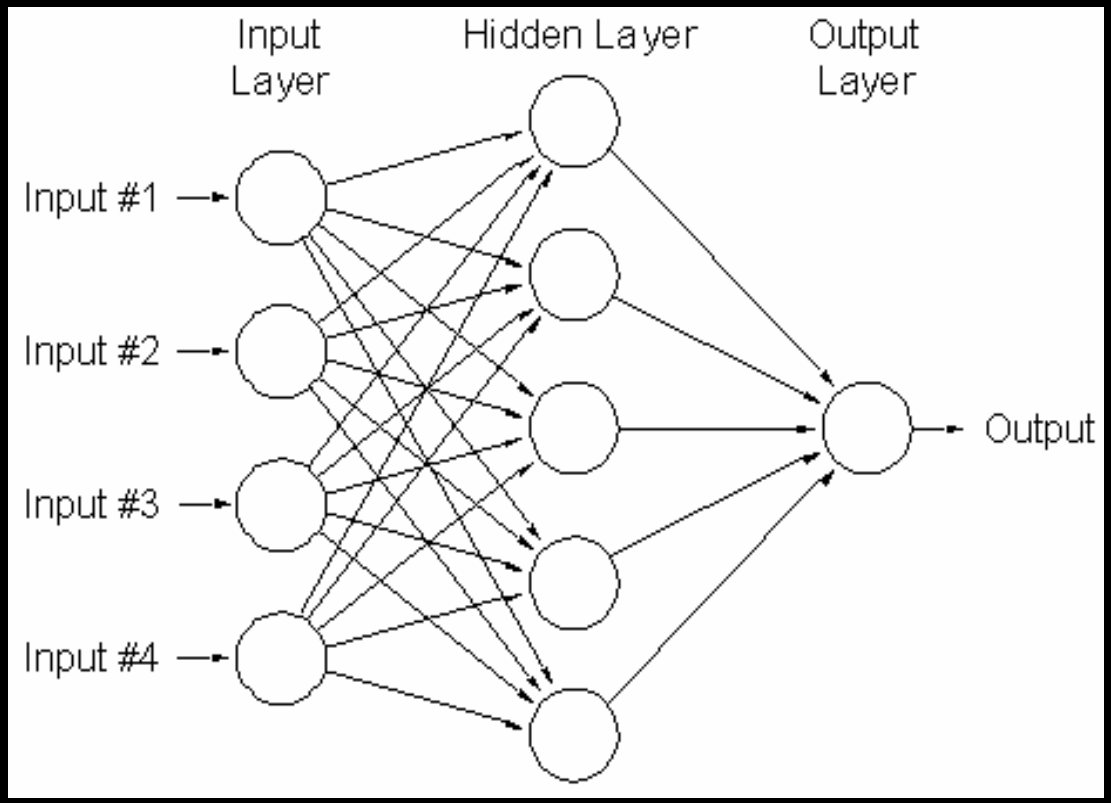
From the Physically Based Models

Such as:

- the Système Hydrologique Européen (SHE)
- the SHETRAN and MIKE-SHE
- the LISFLOOD model
- the TIN-based Real-time Integrated Basin Simulator (tRIBS)
- the TOPographic KInematic wave APproximation and Integration (TOPKAPI)
- etc.



To the ANN Models



Or the DB Mechanistic Models

$$y_t = \frac{\hat{B}(z^{-1})}{\hat{A}(z^{-1})} u_{t-k} + \xi_t$$

Procedure:

- Step 1: classical linear or non-linear black-box model
- Step 2: physical interpretation of resulting model
- Step 3: acceptance if good, parsimonious and physically meaningful
- . . .

Young, P. C., 2001. Data-based mechanistic modelling and validation of rainfall-flow processes. In: *Model Validation: Perspectives in Hydrological Science,* M.G. Anderson and P.D. Bates, (Eds.) Wiley, Chichester, UK. 117-161.

THE CONTEXT
Flood Emergency Management

In general, flood emergency decisions are taken without perfect knowledge of what will happen next in the future.

Operational flood forecasting systems use hydro-meteorological and hydraulic models to provide water stage forecasts that are generally compared with warning or emergency threshold values

PREDICTIVE UNCERTAINTY

The concept of Predictive Uncertainty is not yet well understood, not only by the communities of Meteorologists and Hydrologists, but in particular by the stakeholders, namely the managers of flood emergencies.

In addition, the use of QPF, generally in the form of ensembles, as input to the hydrological models has certainly raised the level of confusion and misunderstanding.

Predictive Uncertainty

In order to understand the meaning of predictive uncertainty, let me pose the following question:

Flooding damages will occur:

(1) when the actual water level overtops the dykes, or,
(2) when the forecasted level overtops the dykes?

The obvious answer is

(1) when the actual future water level overtops the dykes

This answer has a strong implication in the definition of the predictive uncertainty

Predictive uncertainty is obviously the uncertainty that we have on the occurrence of a real future value as for instance the water level in 12 hours from now.
This must not be confused with "model uncertainty".

So what is the role of the forecasting model(s)? Why are we using models?

In clarifying to hydrologists the meaning of predictive uncertainty, Krzysztofowicz (1999), points out that

"*Rational decision making (for flood warning, navigation, or reservoir systems) requires that the total uncertainty about a hydrologic predictand (such as river stage, discharge, or runoff volume) be quantified in terms of a probability distribution, conditional on all available information and knowledge.*"

and that

"*Hydrologic knowledge is typically embodied in a deterministic catchment model*".

In other words….

The decision maker, uncertain on what will happen, tries to use all available information (and what is better than a hydrological forecast?).

but…

The uncertainty he must describe and use in the decision making process is the uncertainty of the future value of the quantity of interest (for instance the water stage), now conditional on the model forecast

and not

The uncertainty of the forecasted quantity, namely the model output.

Predictive Uncertainty
(simplified)

$$f\left(y_{t+n\Delta t}\middle|\mathbf{y}_t,\mathbf{x}_t,\hat{\mathbf{y}}_{t,t+n\Delta t},\hat{\mathbf{x}}_{t,t+n\Delta t}\right)$$

where

$y_{t+n\Delta t}$	is the value that the predictand of interest will take n time steps from now into the future
$\mathbf{y}_t$	is the ensemble of all the measurements of the predictand up to the present time
$\mathbf{x}_t$	is the ensemble of all the measurements of the inputs up to the present time
$\hat{\mathbf{y}}_{t,t+n\Delta t}$	is the model prediction n time steps from now into the future
$\hat{\mathbf{x}}_{t,t+n\Delta t}$	is the ensemble of all the input predictions up to n time steps from now

Model Uncertainty
(simplified)

$$f\left(\hat{y}_{t,t+n\Delta t}\middle|\mathbf{y}_t,\mathbf{x}_t,\hat{\mathbf{x}}_{t,t+n\Delta t}\right)$$

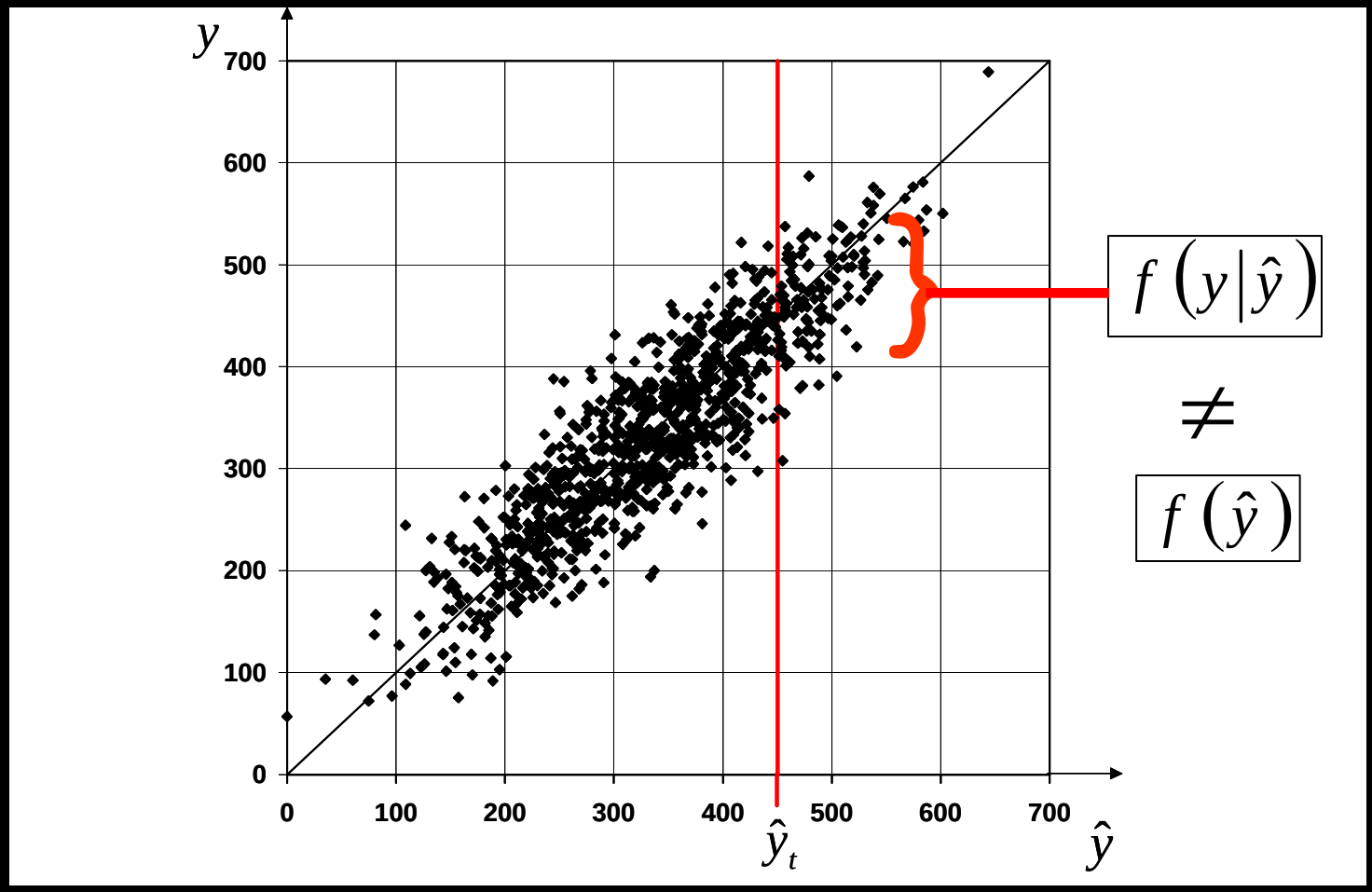
$\hat{y}_{t,t+n\Delta t}$ is the model prediction n time steps from now into the future

$\mathbf{y}_t$ is the ensemble of all the measurements of the predictand up to the present time

$\mathbf{x}_t$ is the ensemble of all the measurements of the inputs up to the present time

$\hat{\mathbf{x}}_{t,t+n\Delta t}$ is the ensemble of all the input predictions up to n time steps from now

Simple representation of conditional uncertainty



y Predictand

$\hat{y}$ Model

The Normal Quantile Transform

In order to use a convenient multivariate distribution, avoiding making hypotheses on the actual distributions of the predictand and the model output, one can use the Normal Quantile Transform

Van der Waerden B.L. (1952). Order tests for two-sample problem and their power I. *Indagationes Mathematicae*, **14**: 453-458.

Van der Waerden B.L. (1953a). Order tests for two-sample problem and their power II. *Indagationes Mathematicae*, **15**: 303-310.

Van der Waerden B.L. (1953b). Order tests for two-sample problem and their power III. *Indagationes Mathematicae*, **15**: 311-316.

Kelly, K. S., and R. Krzysztofowicz, (1997) A bivariate meta-Gaussian density for use in hydrology, *Stochastic Hydrol. Hydraul.*, *11*, 17–31.

The Normal Quantile Transform STEPS

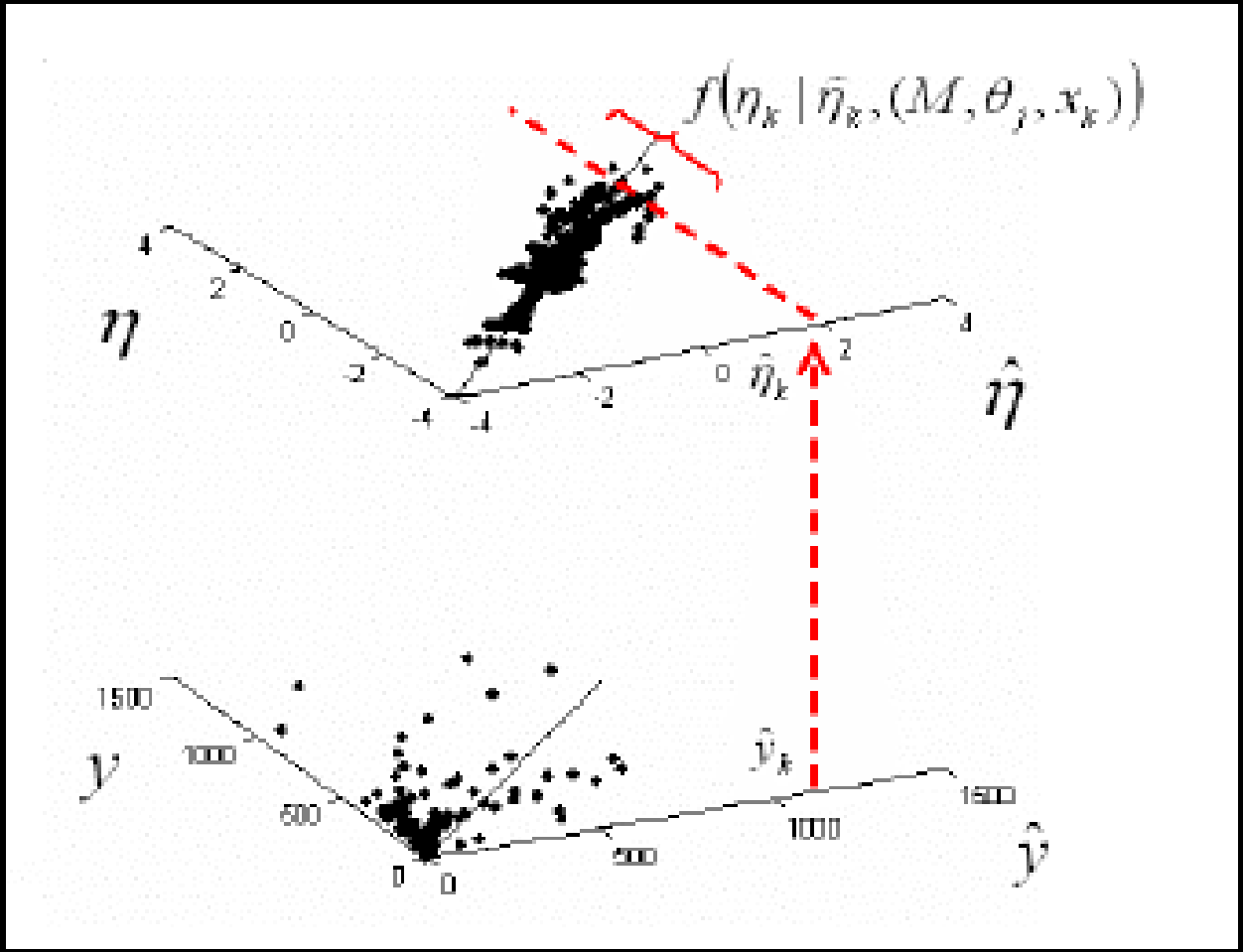
1) Order y_t in ascending order and attach the probability $P_i = i/n+1$ to the i^{th} element of the ranked vector

2) Convert P_i into the Standard Normal variable η_t corresponding to probability P_i

3) Order $\hat{y}_t$ in ascending order and attach the probability $P_j = j/n+1$ to the j^{th} element of the ranked vector

4) Convert P_j into the Standard Normal variable $\hat{\eta}_t$ corresponding to probability P_j

The Normal Quantile Transform

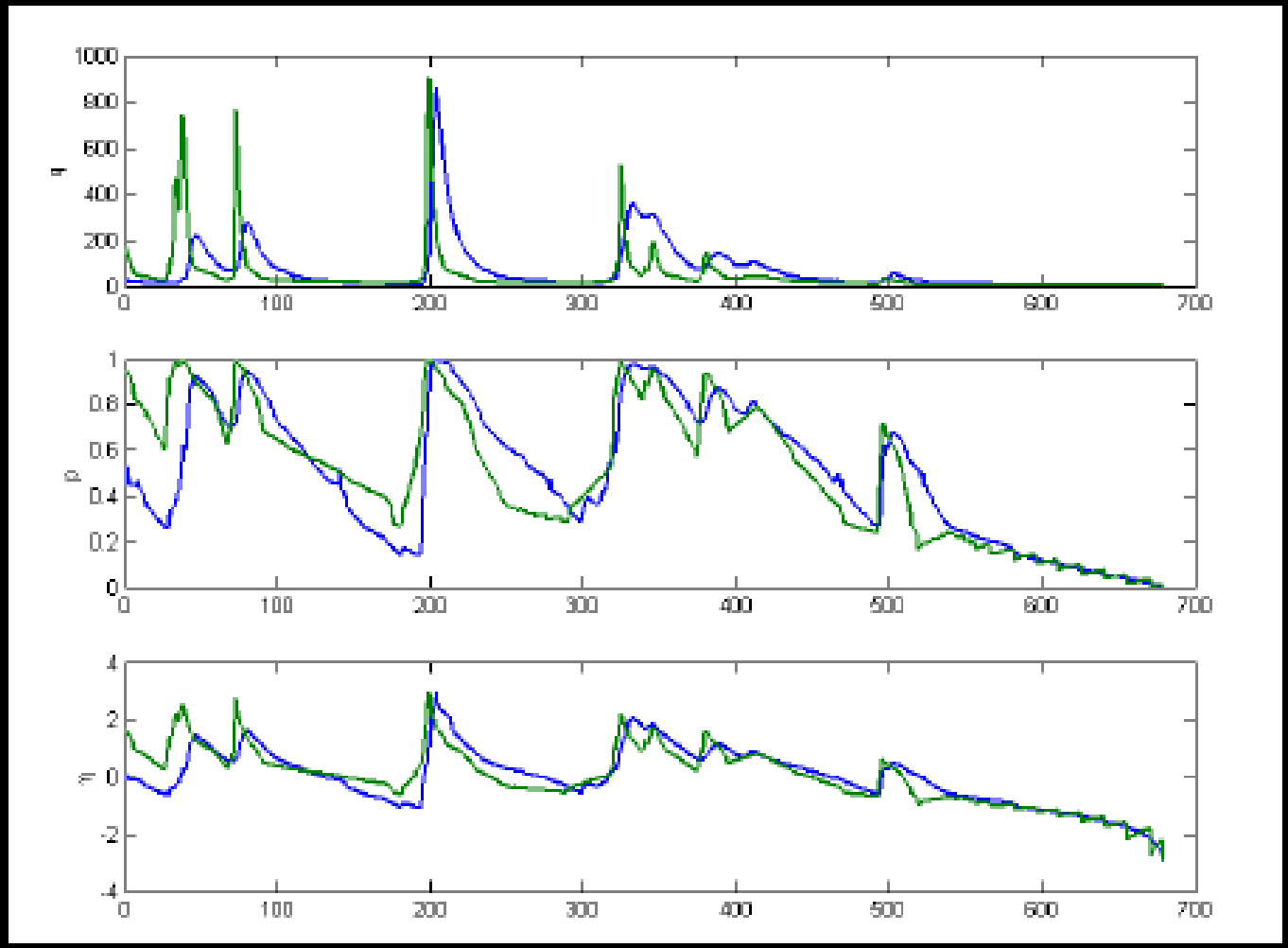


The Normal Quantile Transform

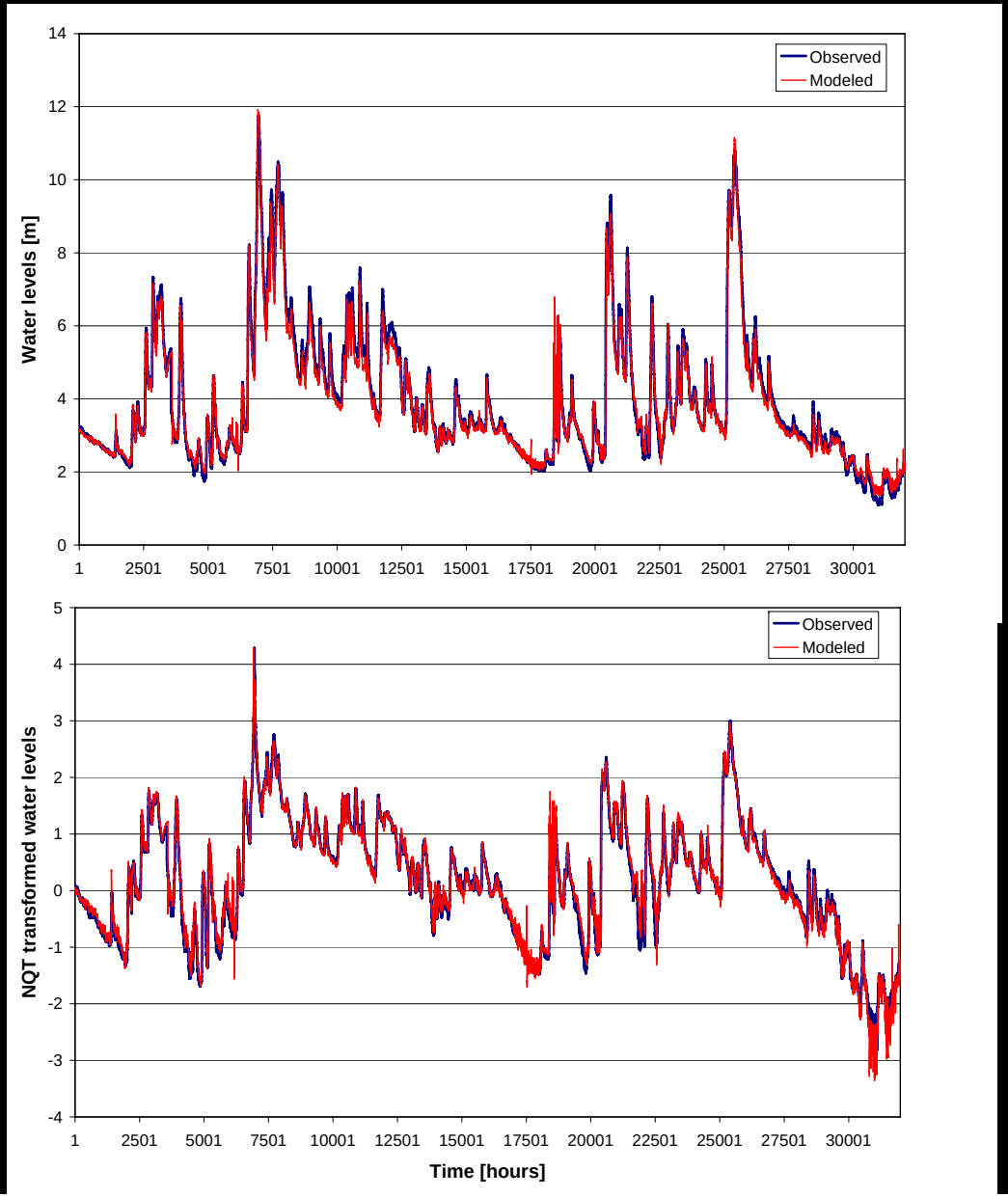
y_t $\hat{y}_t$

P_t $\hat{P}_t$

η_t $\hat{\eta}_t$



The Po river example



Real space

NQT space

NOTE

The NQT does not preserve the Product Moment Correlation, because of the non-linear transformation, but fully preserves the Order Correlation also known as Spearman Rank Correlation

The New Conditional Processor
(Background)

Krzysztofowicz approach has many limitations:

- It uses an auto-regressive model as the a priori model (for instance, this type of model is not suitable for flood routing)

- The a priori model is developed in the transformed space, not in the original one

-The a priori model is implicitly assumed to be independent from the "deterministic" model

- It has a scalar formulation

The New Conditional Processor

If one can make the hypothesis that all the transformed variables follow a multi-Gaussian joint probability density, a more natural approach would be to:

- Develop a set of models in the real untransformed space (one or more than one)

- Build the joint probability density in the Gaussian space (Predictand, a priori model, deterministic model, etc.)

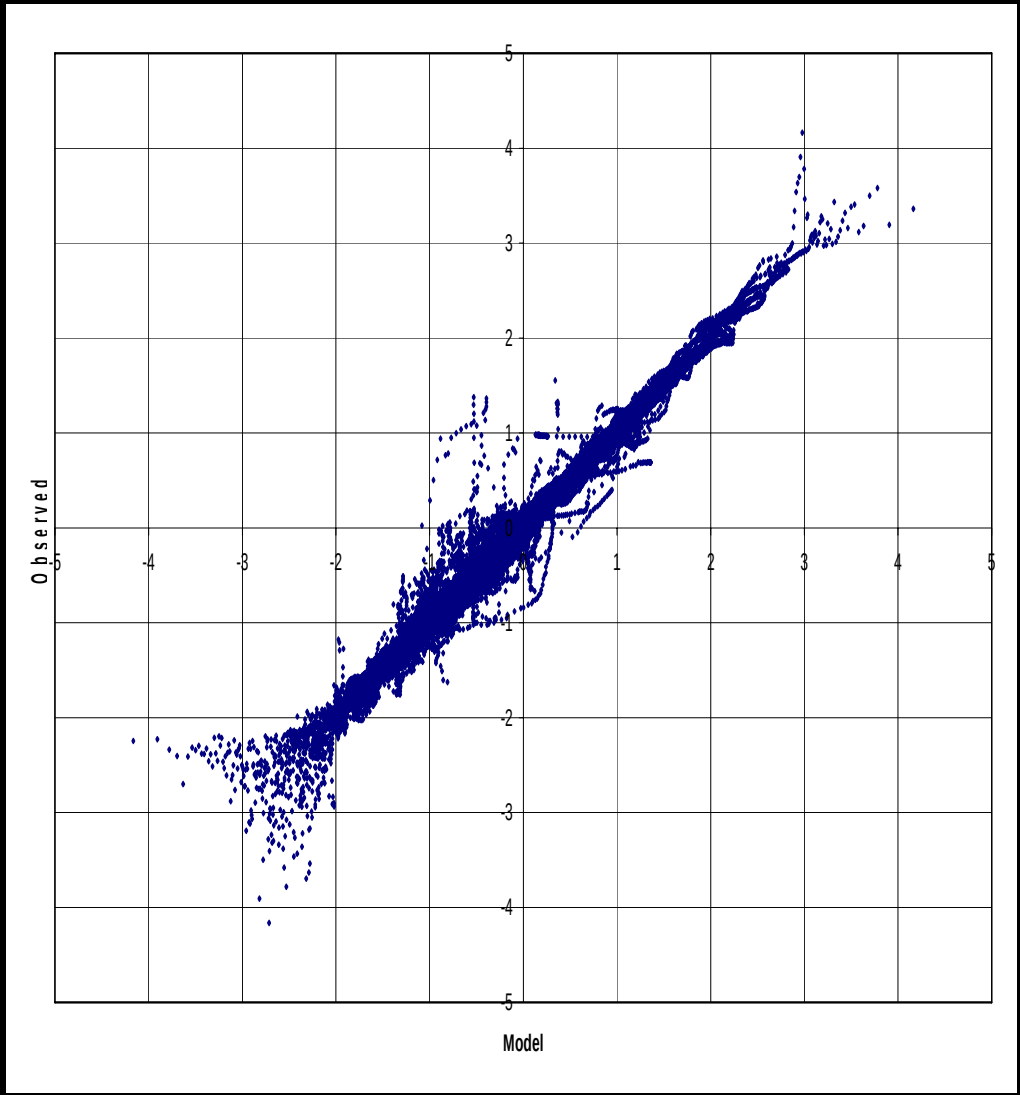
- Simply compute the probability of the predictand conditional on ALL the model predictions

The New Conditional Processor
(Requirements)

The NQT guarantees that the marginal distributions of the transformed variables are all N(0,1).
This does not guarantee that the joint distribution will actually be a multi-Normal, unless one can show that the dependence of the different variables in the transformed space is linear.
This requiement is the same needed by Krzysztofowicz Bayesian processor and it is generally satisfied.

The Po river example

Linear correlation in the Normal space



A Useful Property of the Multivariate Normal Distribution

Given a vector of random variables $\mathbf{x} = \begin{bmatrix} \mathbf{y} \\ \hat{\mathbf{y}} \end{bmatrix}$ Normally distributed with

Mean $\boldsymbol{\mu}_{\mathbf{x}} = \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{y}} \\ \boldsymbol{\mu}_{\hat{\mathbf{y}}} \end{bmatrix}$ and Variance $\boldsymbol{\Sigma}_{\mathbf{xx}} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{yy}} & \boldsymbol{\Sigma}_{\mathbf{y}\hat{\mathbf{y}}} \\ \boldsymbol{\Sigma}_{\hat{\mathbf{y}}\mathbf{y}} & \boldsymbol{\Sigma}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} \end{bmatrix}$ this implies that also

$\mathbf{y} \approx N\left(\boldsymbol{\mu}_{\mathbf{y}}, \boldsymbol{\Sigma}_{\mathbf{yy}}\right)$ and $\hat{\mathbf{y}} \approx N\left(\boldsymbol{\mu}_{\hat{\mathbf{y}}}, \boldsymbol{\Sigma}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}\right)$ are Normally distributed

The conditional distribution of y given yhat is also Normal

$$N\left(\boldsymbol{\mu}_{\mathbf{y}|\hat{\mathbf{y}}}, \boldsymbol{\Sigma}_{\mathbf{yy}|\hat{\mathbf{y}}}\right)$$

with conditional Mean $\boldsymbol{\mu}_{\mathbf{y}|\hat{\mathbf{y}}} = \boldsymbol{\mu}_{\mathbf{y}} + \boldsymbol{\Sigma}_{\mathbf{y}\hat{\mathbf{y}}}\boldsymbol{\Sigma}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}^{-1}\left(\hat{\mathbf{y}} - \boldsymbol{\mu}_{\hat{\mathbf{y}}}\right)$

and conditional Variance $\boldsymbol{\Sigma}_{\mathbf{yy}|\hat{\mathbf{y}}} = \boldsymbol{\Sigma}_{\mathbf{yy}} - \boldsymbol{\Sigma}_{\mathbf{y}\hat{\mathbf{y}}}\boldsymbol{\Sigma}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}^{-1}\boldsymbol{\Sigma}_{\hat{\mathbf{y}}\mathbf{y}}$

Advantage of the proposed approach

- It allows to combine together a wide variety of different models without the need of using the so-called Bayesian Model Averaging

- It allows to have multiple outputs, benefitting from spatial correlation (for instance several water levels along the same river)

- It is a direct approach and it avoids the need for searching optimal weights, as required by the Bayesian Model Averaging

EXAMPLES

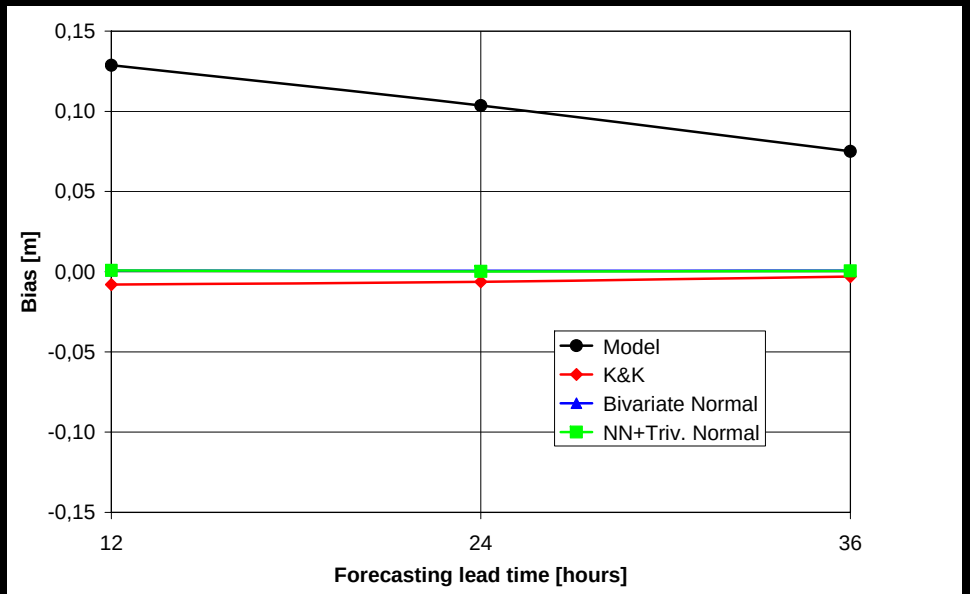
The Po river example

Combination of

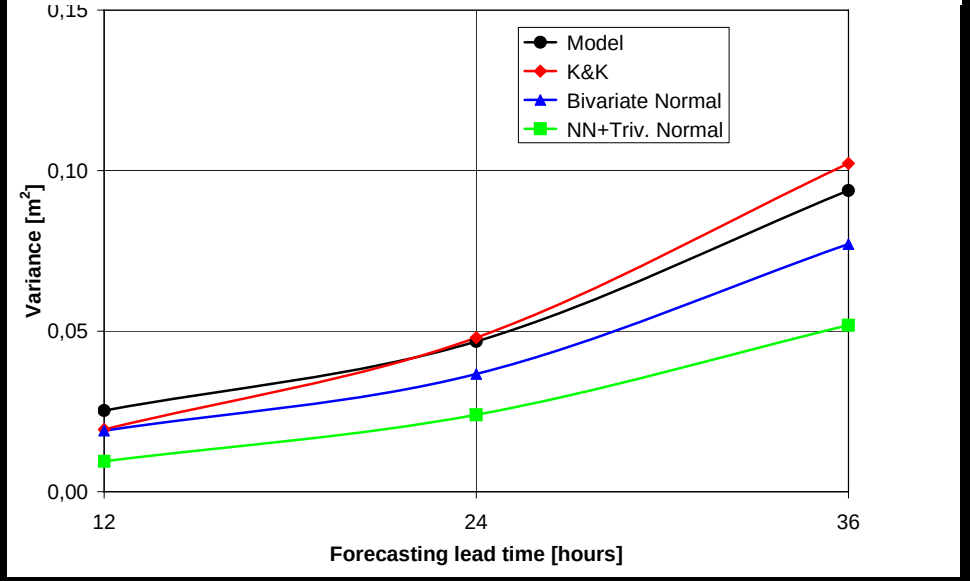
1) A flood routing model

2) A Nearest Neighbour model

The Po river example

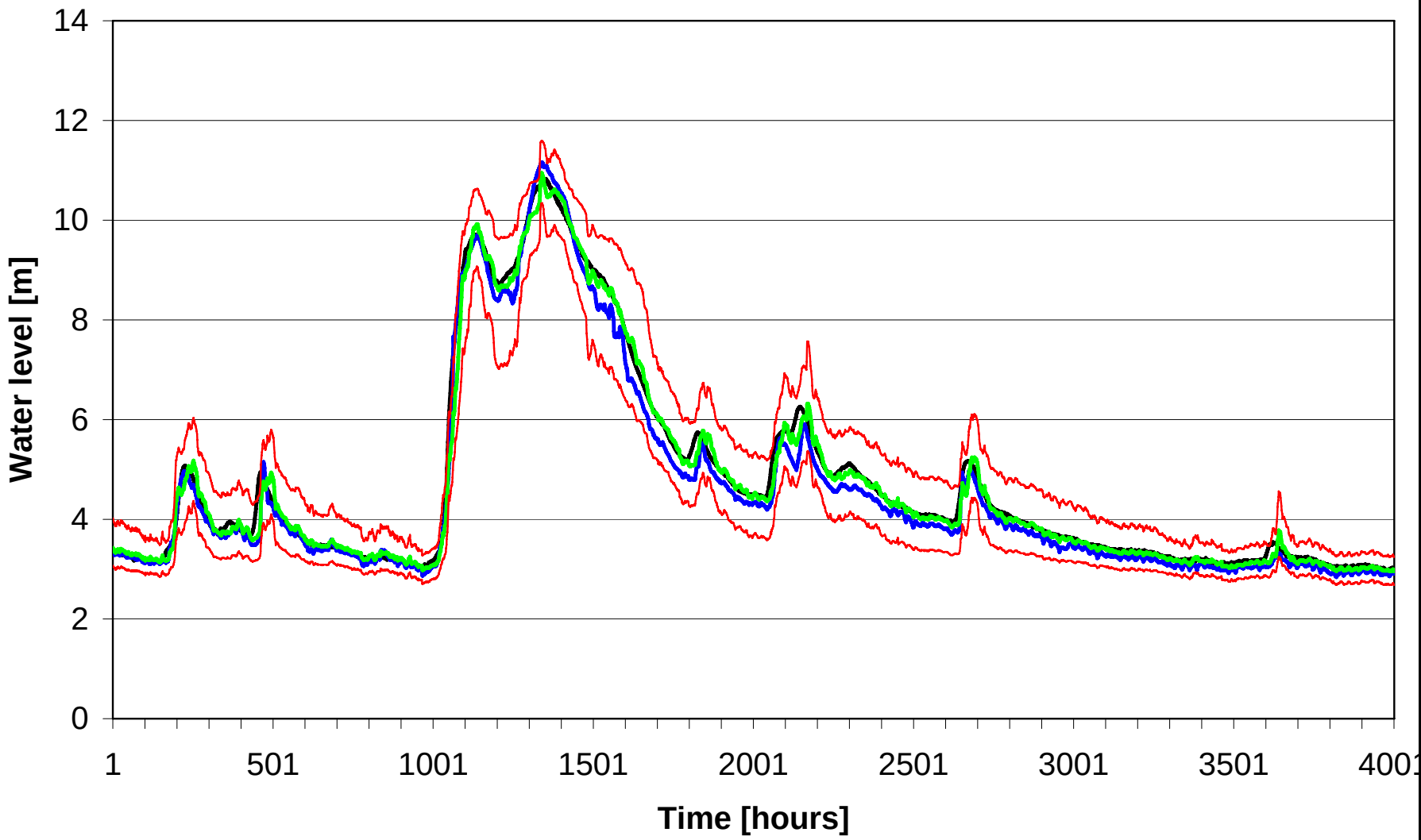


Bias



Error Variance

The Po river example



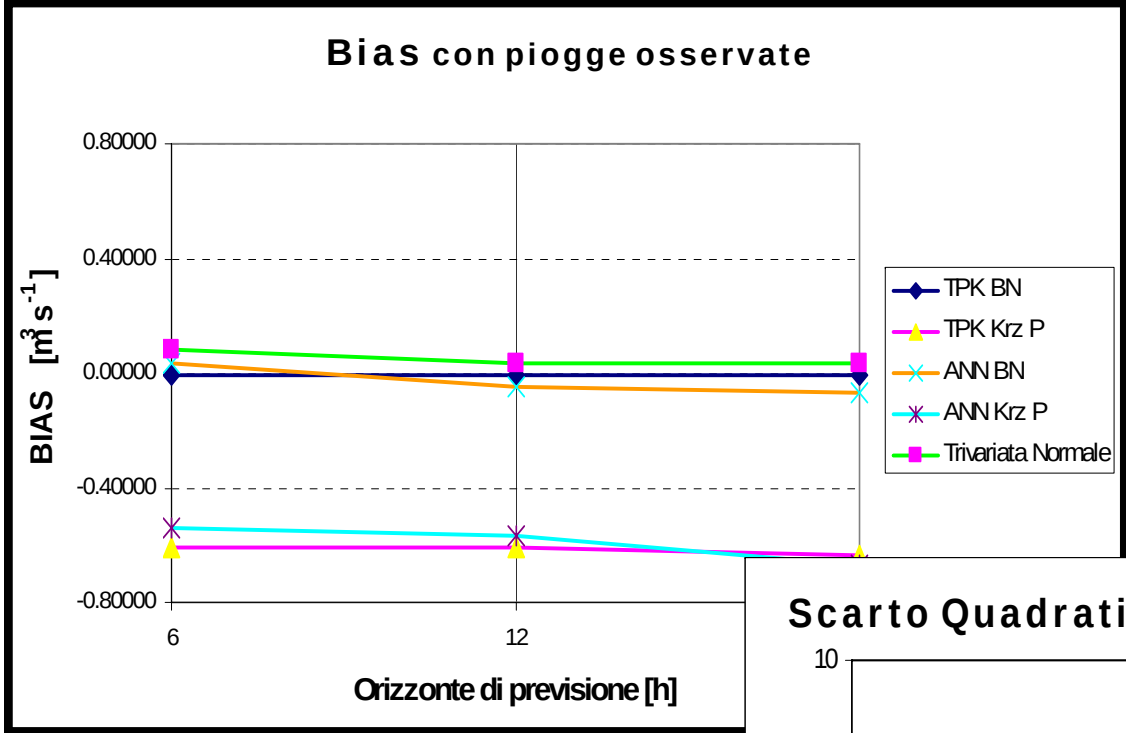
The Parma river example

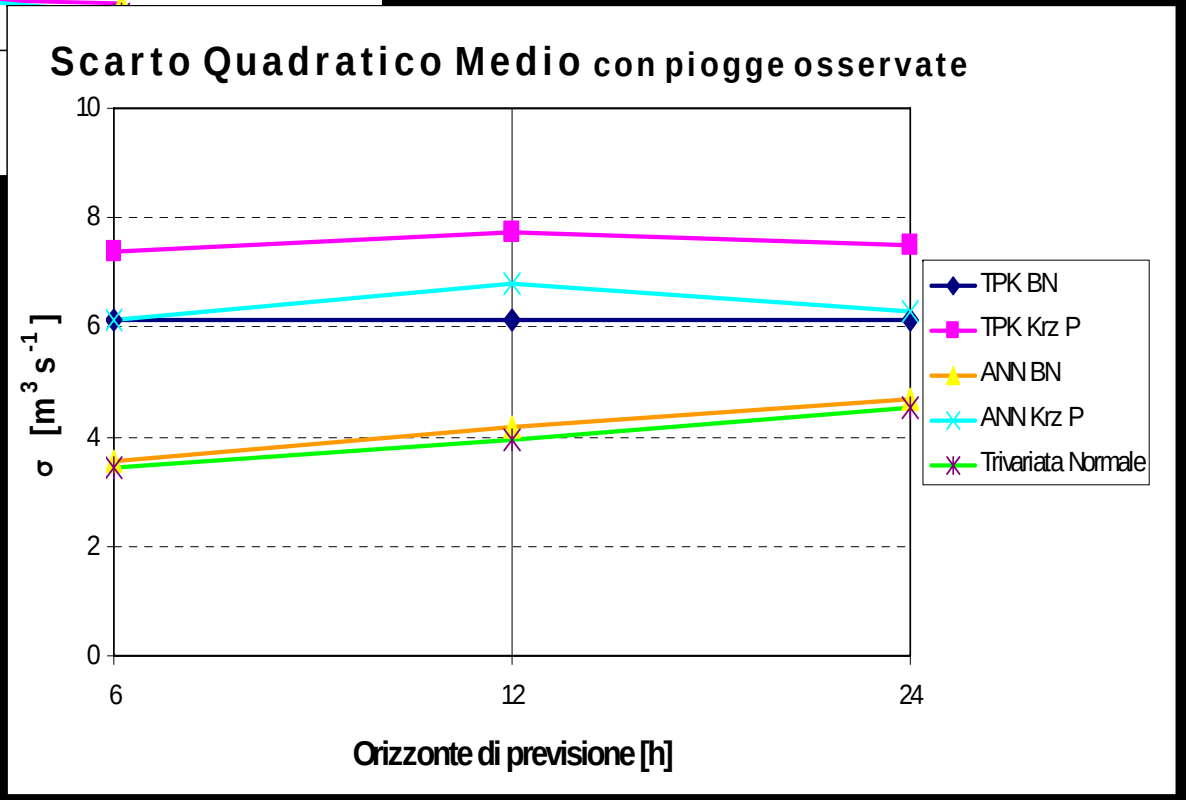
Combination of

1) A rainfall-runoff model (TOPKAPI)

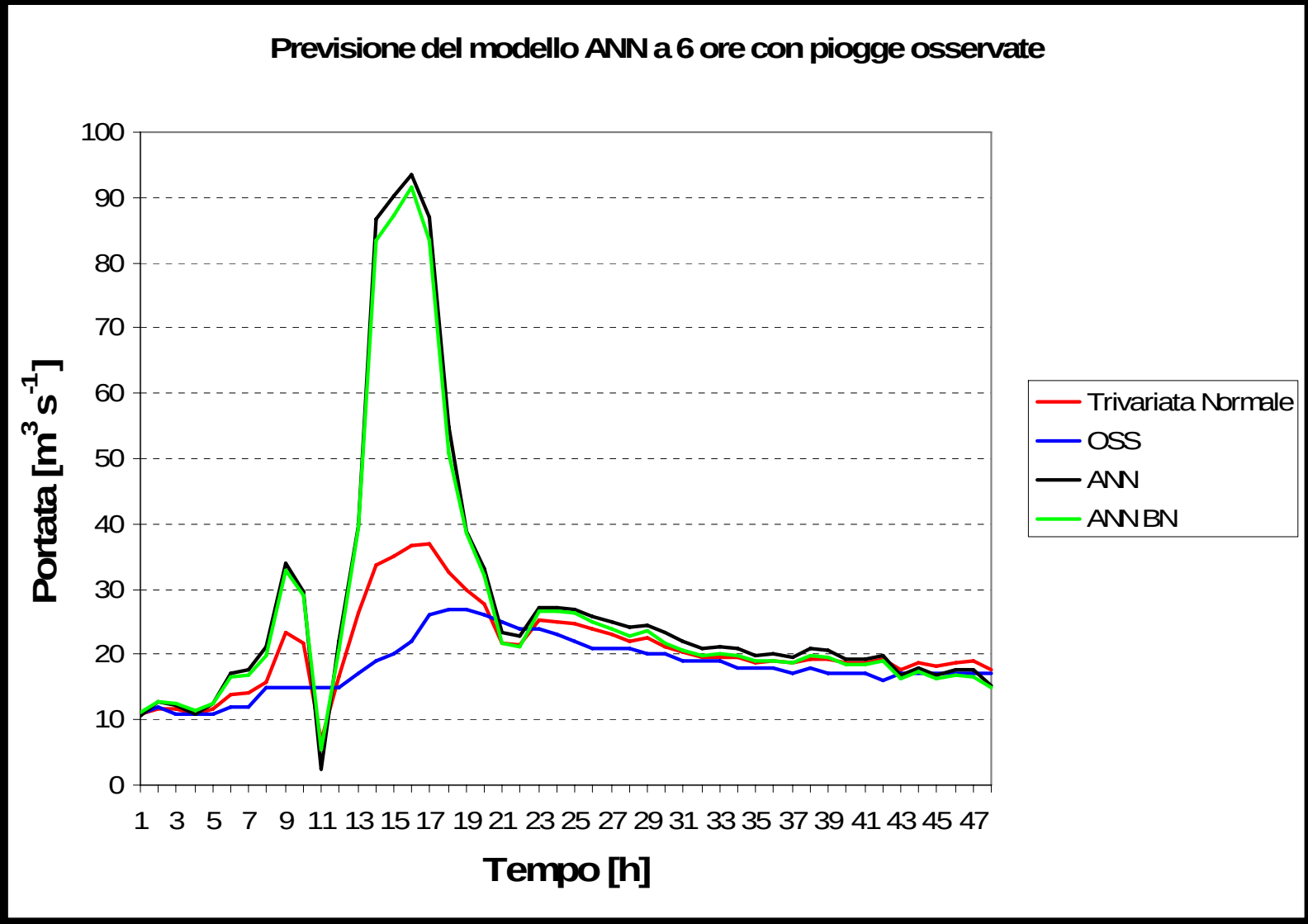
2) An Artificial Neural Network model

The Parma river example





The Parma river example



CONCLUSIONS

The proposed approach opens several interesting perspectives still to be explored.

For instance, the inclusion of input (QPF) Forecasting Uncertainty, can also be taken into account

$$f\left(y_{t+n\Delta t}\middle|\mathbf{y}_t,\mathbf{x}_t,\hat{\mathbf{y}}_{t,t+n\Delta t}\right)=\sum_{i=1}^{m} f^{\,i}\left(y_{t+n\Delta t}\middle|\mathbf{y}_t,\mathbf{x}_t,\hat{\mathbf{y}}^{\,i}_{t,t+n\Delta t},\hat{\mathbf{x}}^{\,i}_{t,t+n\Delta t}\right)\Pr\left(\hat{\mathbf{x}}^{\,i}_{t,t+n\Delta t}\right)$$

$$\approx \frac{1}{m}\sum_{i=1}^{m} f^{\,i}\left(y_{t+n\Delta t}\middle|\mathbf{y}_t,\mathbf{x}_t,\hat{\mathbf{y}}^{\,i}_{t,t+n\Delta t},\hat{\mathbf{x}}^{\,i}_{t,t+n\Delta t}\right)$$

CONCLUSIONS

But most of all, the proposed approach allows to combine models of different nature (deterministic, stochastic, etc.) with the main objective of increasing and incorporating into the forecast, all the available information.

CONCLUSIONS

And, from a hydrological perspective, let us hope that this will finally allow us to reconcile physically based with data driven models to the benefit of the flood managers.

Thank you
for your attention