Quantification and Reduction of Predictive Uncertainty for Sustainable Water Resources Management (Proceedings of Symposium HS2004 at IUGG2007, Perugia, July 2007). IAHS Publ. 313, 2007.

"Outlier" catchments: what can we learn from them in terms of prediction uncertainty in rainfall–runoff modelling?

NICOLAS LE MOINE, VAZKEN ANDREASSIAN, CHARLES PERRIN & CLAUDE MICHEL

Cemagref, Hydrology and Water Quality Research Unit, PB 44, F-92163 Antony cedex, France
nicolas.le-moine@cemagref.fr

Abstract What exactly are the catchments that we usually exclude from our data sets before submitting a paper to a conference, on the grounds that our models fail to represent their behaviour? What can be the consequences of the commonly-used (but rarely discussed) data set cleansing practices: does it help us improve our models? Does it contribute to making our hydrological simulations less uncertain? Or does it just give us a false sense of confidence in our capacity to represent catchment hydrological behaviour? This paper focuses specifically on the "outlier" catchments found in a large set of 1045 French catchments. This large catchment set allows statistical quantification of the likely sources of model failure; it shows regional clustering (linked with the geology), the surprising effect of catchment area (the largest basins get the best performances) and last, that noise in input data is in no way sufficient to explain the difficulties of five rainfall–runoff models in representing catchment behaviour.

Key words rainfall–runoff modelling; outliers; GR4J

INTRODUCTION

Who would care to publish a "failure story"?

When a modeller successfully reproduces the behaviour of a given catchment with his model, he will publish his results in the form of a scientific paper in a peer-reviewed journal. In the case of failure, he will probably give up its study or discard the catchment as an "outlier". Obviously, the publication of "success stories" is on the basis of scientific recognition, and this has precluded the publication of "failure stories", while the corresponding catchments are treated as pariahs. Anyway, who would care to publish a failure story? Who would dare to offer to their peers such an easy case for rejecting the paper?

But what are exactly these outlier catchments? Where are they located? Why are our models unable to reproduce their behaviour? This question has rarely been asked, probably because it requires large catchment sets, and because only a few research groups working on generic models and on regionalization issues have been able to gather these catchment sets (see on this topic, Andréassian *et al*., 2006). When working on large catchment sets, hydrologists will necessarily (statistically) encounter catchments which refuse to comply to the proposed model structure: should these catchments remain hidden?

Is it not too easy to impute all model failures to data quality problems?

To our knowledge, Boughton (2006) is one of the only authors to have published a study where he specifically addresses the issue of problematic catchments (catchments that he had preferred to set apart in a previous analysis). However, he has a preset interpretation of model failures, which he attributes to data problems. Indeed, he states that: "*if given good quality data, any of the modern water balance models will give good quality results*, and that *the results from rainfall–runoff modelling are more dependent on the quality of the input data than on the model*". Such a statement may be partly true, but it is extremely difficult to verify, since it would require a neutral data screening procedure, i.e. one that is independent from the type of model to be tested. For example, if we want to select catchments for the assessment of a rainfall–runoff (RR) model, we should never use a procedure where concepts related to the RR transformation intervene, otherwise we will enter into a form of circular reasoning, where we judge of the suitability of our model according to data which have been previously found compliant with this very model.

On the topic of data quality impact on model behaviour, Oudin *et al.* (2006b) made a fairly comprehensive study showing how systematic and random errors in precipitation and potential evapotranspiration input do reduce model performance. But the authors did not attempt to judge the initial error content of the measured (uncorrupted) series, since it was an inextricable combination of model error and input error.

We do not want here to underestimate the tremendous difficulty of measuring hydrometeorological variables. However, it is obviously excessive to act as if only problematic data could justify problematic simulations, and as if nothing could be improved in RR models. We must stop hiding these problematic catchments if we want to progress in our representation of catchment behaviour. Let us give a quick example: if a catchment yields more runoff that it yields rainfall, it will probably be discarded on the grounds that "areal rainfall is in error". Indeed, the closure of the water balance is often used as a criterion to discard apparently bad data sets. However, on many catchments, there may be intercatchment groundwater flows to or from deep aquifers that may explain why the water balance cannot be closed. In these cases, the data should not be blamed; however, the model should be, for not attempting to detect and represent these possible groundwater exchanges.

A STUDY FOCUSING ON PROBLEMATIC CATCHMENTS

In this paper, we use several rainfall–runoff models and a large data set of 1045 French catchments. We calibrate each model on each of the catchments, and then control it on a different time period. We then focus specifically on the catchments which obtain the poorest results (i.e. whose control efficiency is less than an arbitrary threshold of 60%), and try to understand the hydrological determinants of poor efficiency: why do these catchments "refuse" to be modelled? Conversely, why do our models fail to represent the behaviour of these catchments?

After presenting the large catchment set on which our analysis hinges, we will organize our study of problematic catchments while trying to answer to the following questions: Is there some geographic clustering among these catchments? Is there some

form of scale logic (does catchment size matter)? Can the noise in the input data be held mainly responsible for catchment modelling problems?

MATERIAL

The GR4J rainfall–runoff model

The model used here is GR4J, a parsimonious daily lumped continuous rainfall–runoff model with four parameters to calibrate (Perrin *et al*., 2003). Parameters (see Table 1) were calibrated using a local search procedure.

Table 1 List of the parameters of the GR4J rainfall–runoff model and their meaning.

X1	Capacity of the production reservoir (mm)
X2	Intercatchment groundwater flow magnitude (mm)
X3	Capacity of the nonlinear routing reservoir (mm)
X4	Unit hydrograph time base (day)

Alternative model structures

We also use modified versions of four well-known RR model structures (TOPMO-derived from TOPMODEL, HBV, IHACRES and SMAR). A detailed description is outside of the scope of this paper, and we will just state here that these models are widely-used and that they have been shown by their authors and in the comparison of Perrin *et al*. (2001) to yield good results over a variety of catchment types (refer to Perrin (2000) for a complete description of model structures).

Objective function

The objective function (*OF*) is the Nash & Sutcliffe (1970) criterion computed on root square transformed flows ($\sqrt{Q}$), which was shown to be a good compromise between several alternative criteria (Oudin *et al*., 2006a):

$$OF(\%) = 100\left\{1 - \frac{\sum_j \left(\sqrt{\hat{Q}_j} - \sqrt{Q_j}\right)^2}{\sum_j \left(\sqrt{Q_j} - \overline{\sqrt{Q}}\right)^2}\right\} \quad (1)$$

in which the summation is made on all the time steps j of the period where observed Q_j and $\hat{Q}_j$ computed flows are available.

Catchment set

We used a database of 1045 French catchments located throughout France (see Fig. 1) with daily rainfall, runoff and Penman PE data over the 1995–2005 time period. Catchments where snow has a significant hydrological role were excluded, because the versions of the RR models used here did not have a snowmelt module.

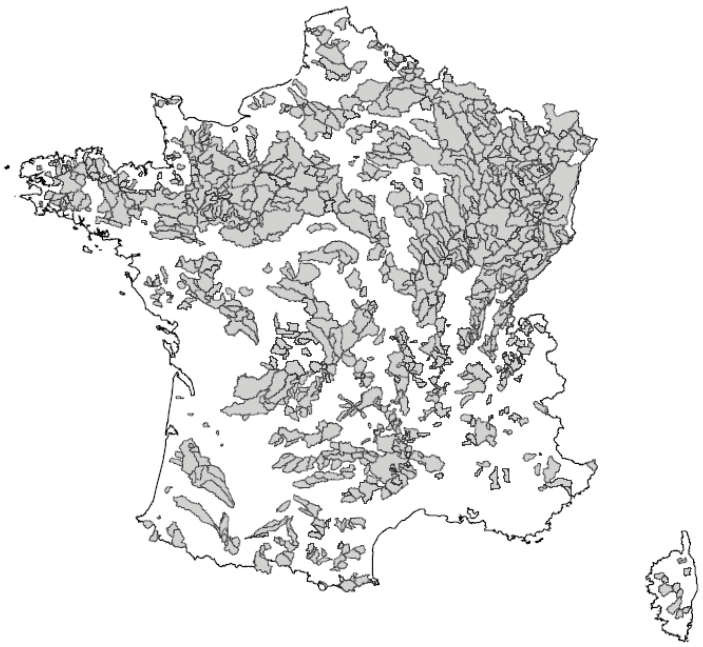

Fig. 1 Location of the 1045 basins used in this study.

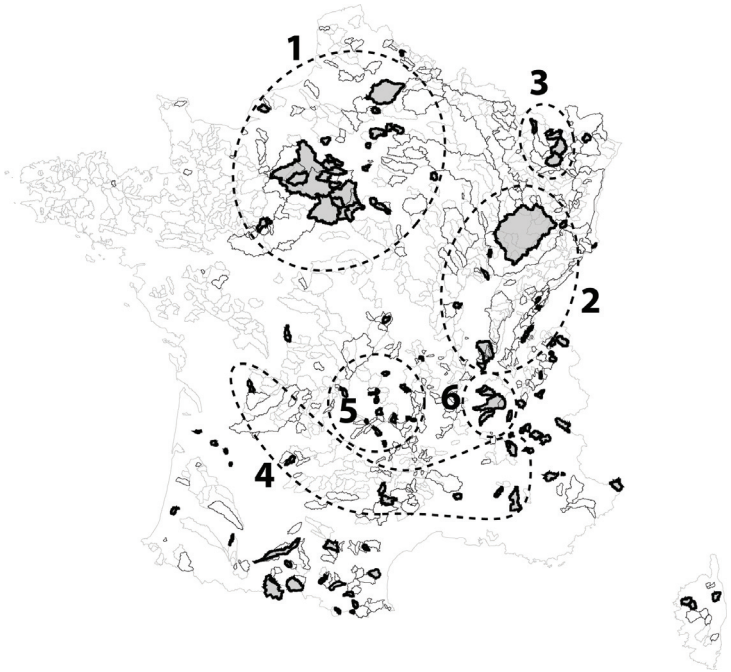

Fig. 2 Map of NS efficiencies of the GR4J RR model in control mode. The thin, light grey outlines are the catchments where the model performs well (NS > 80%). The catchments with a medium efficiency (60% < NS < 80%) are shown with a thin black outline. The shapes filled in grey with a thick black outline are outlier catchments where NS < 60%.

Geography of "outlier" catchments

Figure 2 shows the map of NS efficiencies (in validation mode) for the GR4J model. Our first observation is that the outlier catchments seem to be clustered, but without

any particular correlation with size, elevation, or precipitation regime (many of them being located in regions of plains and oceanic precipitation regime). However, several well-known groups of catchments can be identified on this map:

Cluster 1: catchments of the Chalk in the Paris basin. These catchments have a very complex hydrological functioning, with a strong correlation to the large, fractured aquifer of the Chalk.

Cluster 2: in this group we find many catchments with karstic influences, located in Jurassic limestone (Jura mountain range and Burgundy).

Cluster 3: gathers a couple of eastern tributaries of the Moselle River known for their particular hydrological functioning (wetlands and salty exurgences of the Vosges sandstone aquifer) (exsurgence = diffuse (nonpoint souce) groundwater outflow).

Cluster 4: catchments in cluster 4 are similar to those of cluster 2, in regions of karstic plateaus to the south and west of the Massif Central highlands (Perigord, Causses) in which are found the largest karstic springs in France, such as the Fontaine de Vaucluse or the Touvre River.

Cluster 5: the model fails to reproduce the behaviour of several catchments influenced by volcanic aquifers in the Massif Central highlands.

Cluster 6: is an interesting example of the strong, periodic influence of groundwater on streamflow. This group of eastern tributaries of the Rhône (Vega, Vesonne, Gère, Collières, Herbasse) are subject to a phenomenon called "zulin" or "julin", which is a periodic exsurgence flow from the molasse (sandstones) aquifer.

Are larger catchments more difficult to model?

As the models used here are lumped ones, many hydrologists would expect that the largest catchments are the most difficult to model. To verify this common-place idea, we draw in Fig. 3(a) the plot of NS efficiency (in control) *versus* catchment area. It seems that there are far fewer model failures for larger catchments than for medium and small ones. In Fig. 3(b) we divided the catchment set into three classes of area and plotted the distributions of NS efficiencies in control for these three classes (so that the result is not influenced by the number of catchments in each class). We can see clearly that, on average, the larger the catchment, the higher the performance.

What model variability could be expected from data errors?

One of the main difficulties with outlier catchments lies in resisting the temptation to attribute all the problems to data errors. The fact is that data errors do exist but that on a catchment-by-catchment basis, it is impossible to judge impartially (i.e. independently from one or the other model) the *a priori* quality of a data set. In this section, we present an approach which will help us judge whether the variability found in our results can be explained solely by data errors. To this aim, we used the RR models mentioned previously (GR4J, TOPMO, HBV, IHACRES and SMAR) in order to:

– estimate the variability of model efficiency among well-known and widely-used models; and

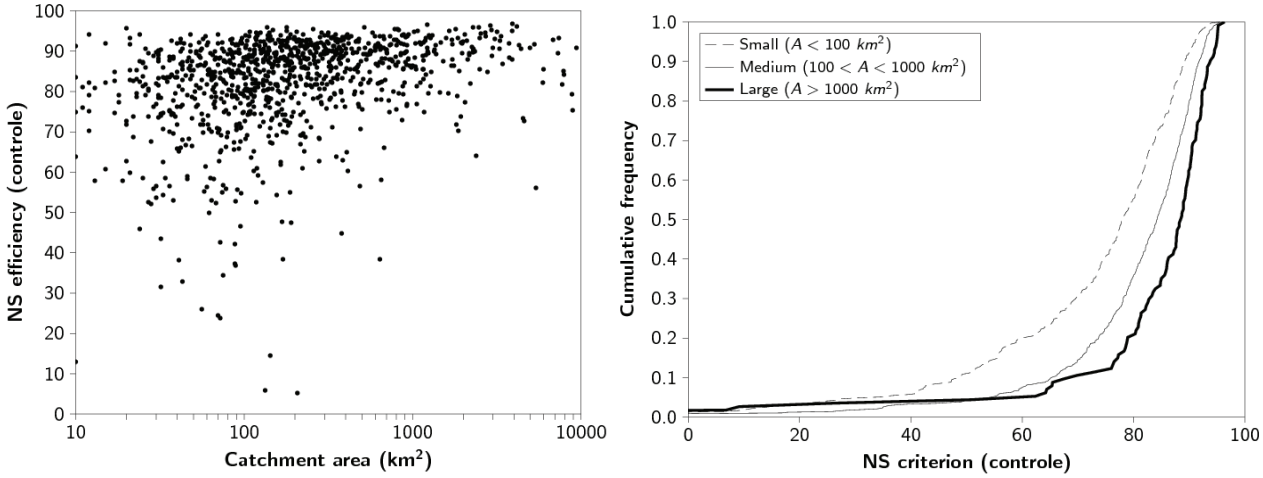

Fig. 3 GR4J performances (Nash-Sutcliffe criterion value in control mode) for: (a) individual values (each point = one catchment); and (b) distribution for three classes of catchments.

– compare the performances of these real models with that of a "perfect" model assessed against imperfect rainfall and runoff data.

METHOD

Let us suppose we have access to the perfect RR model which, given the perfect rainfall P_{true}, would give us the exact discharge $Q_{true} = \hat{Q}(P_{true})$. Lets us now suppose that we feed this model with real, imperfect rainfall P_{mes} and that we compare the results with a real, imperfectly measured discharge series Q_{mes}, different from the "true"' discharge Q_{true}.

In this case, the formulation of the *NS* criterion can be written as:

$$NS = 1 - \frac{\sum_j \left(\hat{Q}_j(P_{mes}) - Q_{j,obs}\right)^2}{\sum_j \left(Q_{j,obs} - \bar{Q}_{obs}\right)^2} = 1 - \frac{\sum_j \left(\hat{Q}_j(P_{mes}) - Q_{j,true} + Q_{j,true} - Q_{j,obs}\right)^2}{\sum_j \left(Q_{j,obs} - \bar{Q}_{obs}\right)^2} \quad (2)$$

(Note that the reference is the observed, imperfect discharge).

The term $\varepsilon_P = \hat{Q}(P_{mes}) - Q_{j,true}$ is the error in discharge due to imperfect rainfall estimates, whereas the term $\varepsilon_Q = Q_{j,true} - Q_{j,obs}$ is an artefact in model error due to discharge measurement errors.

In order to estimate the shape of the distribution of the *NS* criterion for such a perfect model running with imperfect data, we conducted data corruption tests similar to those reported by Oudin *et al.* (2006): they ran a rainfall–runoff model with corrupted data series and compared the simulated discharge with synthetic discharge series given by the same model run with uncorrupted data.

Here we used the following corrupting algorithms:

– For daily rainfall: $P_j^* = (1 + \mu + \sigma \cdot \varepsilon_j) P_j$ where ε_j is a Gaussian, centred noise. This corruption is a mix of a systematic bias and a randomly distributed error expressed as a fraction of rainfall.
– For discharge:

$$Q_j^* = \overline{Q}\left(\frac{Q_j}{\overline{Q}}\right)^{1+\beta}$$

where $\overline{Q}$ is the mean daily discharge. This corruption reflects the systematic bias in low and high flow estimates when extrapolating the rating curve.

Figure 4 shows the resulting *NS* criteria distributions when sampling μ, σ and β in normal distributions to produce either "moderate" or a "strong" noise. For the moderate noise case we take $\mu \sim N(0,0.1)$, $\sigma \sim N(0.2,0.1)$, and $\beta \sim N(0,0.3)$. For the strong noise case, we use $\mu \sim N(0,0.2)$, $\sigma \sim N(0.2,0.1)$, and $\beta \sim N(0,0.4)$. The treatment is applied to all the catchments; each time 10 sets $<\mu, \sigma, \beta>$ are drawn without recalibration of the model.

For comparison, we display on the same figure the distributions of NS efficiencies for the different RR models (GR4J, HBV, IHACRES, SMAR and TOPMO). The bold lines are the distributions of minimal and maximal efficiencies, i.e. the distribution of efficiencies obtained by selecting the best (and the worst) model for each catchment in control.

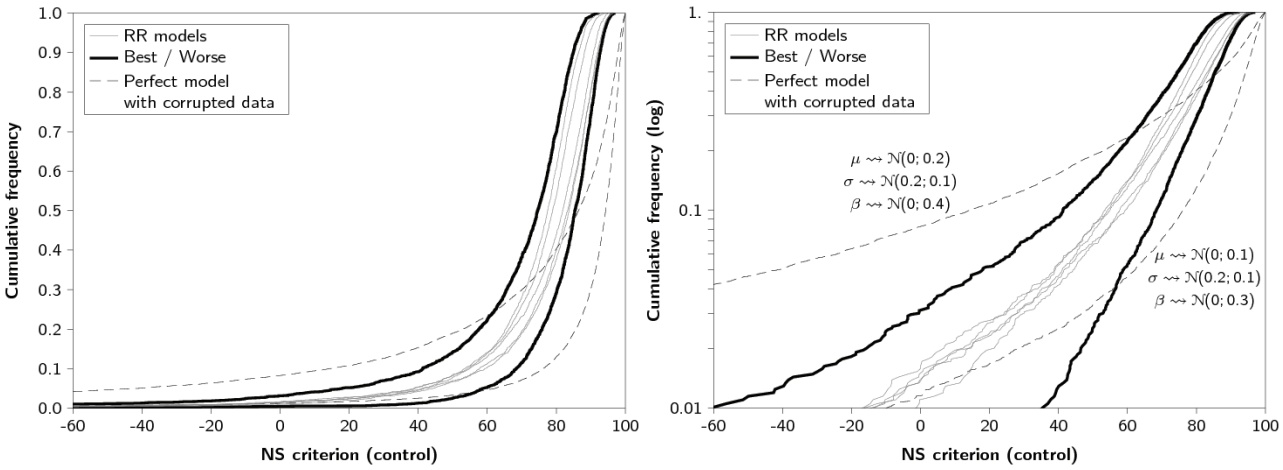

Fig. 4 Distribution of model performances in control mode: comparison of real model distributions with distributions obtained by corrupting synthetic data. Graph (b) is the same as (a) but with a log-axis for the CFD in order to see the tail of the distribution.

The initial conclusions that we can draw from this simple test are the following:

– The distribution of model performance obtained with real data shows that if each individual modeller was to reject the catchments where his model fails (NS in control < 60%), this would result in the removal of more than 20% of the data set. In our view, this would be extremely dangerous.

– If we were to reject a catchment only when all five existing RR models failed, the percentage of discarded catchment would immediately drop from 20% to 5%. Obviously, this would be a better approach. However it would still not be proof for the good quality of the remaining data since model parameters can compensate for measurement errors (Andréassian *et al*., 2001, 2004).
– Last, it is very interesting to note that the theoretical NS distributions (obtained with a "perfect" model fed by corrupted inputs) are very different from the distributions observed for the real RR model. At the median of the distribution, even the worst of the cases ("strong" noise) is superior to the best real RR model: this shows that the imperfection of models has a much greater impact than input noise.

Our conclusion is that any criterion used to evaluate the models (e.g. the NS criterion) will be an inextricable combination of model and data error. But we have to admit that we cannot have access to the "true" states of natural systems: at some level, someone should really criticize observations and data, but this should not be the modellers' responsibility.

CONCLUSION AND PERSPECTIVES

The aim of this paper was to analyse the outliers found in a large set of 1045 French catchments, in order to obtain an overall view of the likely sources of model failure. The results show that:
– the geographical clustering of outliers excludes random measurement errors: it is more likely that regional hydrological difficulties (such as the occurrence of regional groundwater flows) justify the poor performance of the model. Thus, instead of rejecting the catchments, specific work should be done to try to improve the model so that it would represent reality better.
– there is an impact of catchment size on model performance, but it is the opposite of the expected one (larger basins get the best efficiency, even with lumped models!);
– results obtained with synthetic data clearly show that the possible noise in data is not enough to explain the distribution of model performances.

Further work is needed with outlier catchments, in order to generalize this paper's observations. Much can be learnt from model failures and, consequently, their publication should be encouraged.

REFERENCES

Andréassian, V., Hall, A., Chahinian, N. & Schaake, J. (eds) (2006) *Large Sample Basin Experiments for Hydrological Model Parameterization: Results of the Model Parameter Experiment - MOPEX*. IAHS Publ. 307. IAHS Press, Wallingford, UK.

Andréassian, V., Perrin, C. & Michel, C. (2004) Impact of imperfect potential evapotranspiration knowledge on the efficiency and parameters of watershed models. *J. Hydrol.* **286**(1-4), 19–35.

Andréassian, V., Perrin, C., Michel, C., Usart-Sanchez, I. & Lavabre, J. (2001) Impact of imperfect rainfall knowledge on the efficiency and the parameters of watershed models. *J. Hydrol.* **250**(1-4), 206–223.

Boughton, W. (2006) Calibrations of a daily rainfall–runoff model with poor quality data. *Environmental Modelling and Software* **21**, 1114–1128.

Nash, J. E. & Sutcliffe, J. V. (1970) River flow forecasting through conceptual models. Part I - A discussion of principles. *J. Hydrol.* **10**, 282–290.

Oudin, L., Andréassian, V., Mathevet, T., Perrin, C. & Michel, C. (2006a) Dynamic averaging of rainfall–runoff model simulations from complementary model parameterization. *Water Resour. Res.* **42**(7). W07410, doi:10.1029/2005WR004636.

Oudin, L., Perrin, C., Mathevet, T., Andréassian, V. & Michel, C. (2006b) Impact of biased and randomly corrupted inputs on the efficiency and the parameters of watershed models. *J. Hydrol.* **320**, 62–83.

Perrin, C. (2000) Vers une amélioration d'un modèle global pluie–débit au travers d'une approche comparative. PhD Thesis, INPG, Grenoble, France.

Perrin, C., Michel, C. & Andréassian, V. (2001) Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. *J. Hydrol.* **242**, 275–301.

Perrin, C., Michel, C. & Andréassian, V. (2003) Improvement of a parsimonious model for streamflow simulation. *J. Hydrol.* **279**, 275–289.