Upper and lower benchmarks in hydrological modeling

Jan Seibert[1,2], Marc Vis[1], Elizabeth Lewis[3], Ilja van Meerveld[1]

[1] Department of Geography, University of Zurich, Zurich, Switzerland

[2] Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Uppsala, Sweden

[3] School of Engineering, Newcastle University, Newcastle, United Kingdom

Introduction

When assessing the performance of a hydrological model, a question that can be raised is, how good is really good? Despite several calls to use benchmarks (Seibert, 2001; Schaefli and Gupta, 2007; Pappenberger *et al.*, 2014), model performance in the scientific literature, conference presentations and discussions among hydrological modelers is still often solely judged based on the value of some performance measure. For instance a model is rated as well-performing because model efficiency (Nash and Sutcliffe, 1970) values are above 0.7. Some authors (e.g., Moriasi *et al.*, 2007; Ritter and Muñoz-Carpena, 2013) even suggest performance classes based on model efficiency values. Based on our experiences with the application of hydrological models for catchments with largely varying characteristics, we argue that such judgments on model performance can only be made if model performances are related to benchmarks that represent what could and should be expected.

The idea of using benchmarks is by no means new and actually the most commonly used performance measure in hydrological modeling, the model efficiency or Nash-Sutcliffe efficiency (Nash and Sutcliffe, 1970), can be interpreted as the comparison of model simulations with a constant streamflow equal to the observed mean streamflow (lower benchmark) and a perfect fit (upper benchmark). Obviously, this lower benchmark is not too hard to beat, whereas this upper benchmark is hardly achievable in practice. To better evaluate how good model simulations are, more informative lower benchmarks have been suggested (Garrick *et al.*, 1978; Seibert, 2001; Schaefli and Gupta, 2007). However, the use of benchmarks that are taking into account what is possible with the data, i.e. what could and should be expected, is still not common practice in hydrological modeling.

In hydrological modeling, it is never possible to obtain a perfect model fit. This is partly due to the complexity of processes in nature, but also due to errors in observations of the driving data and streamflow. Therefore, the upper benchmark should not be an unrealistic perfect simulation but take potential errors in the data into account. On the other hand, there is usually also a lower limit on how bad a model can be, simply because the driving data ensure that the simulated streamflow cannot be totally off as long as the model respects the basic water balance. We therefore argue that we should evaluate model performance relative to what could be possible and what should be expected. The upper benchmark is an evaluation of what is possible for a certain data set and the lower benchmark is an evaluation of what should be expected for that data set.

Here we make a case for the use of upper and lower benchmarks and suggest possibilities for concrete benchmarks based on simulations using simple hydrological models which implicitly take observation uncertainties in both input and output time series into account. We evaluated how much the upper and lower benchmarks vary for catchments in the UK and the US, and demonstrate the potential use of upper and lower benchmarks to evaluate model performance based on the example of the uncalibrated application of the SHETRAN model to 306 catchments in the UK.

Benchmarks derived from a simple bucket type model

The basic idea of the proposed upper and lower benchmarks is to allow comparison of simulated streamflow time series from any model with those obtained using a simple bucket-type modeling approach. The HBV model (Bergström, 1976; Lindström *et al.*, 1997; Seibert, 1999) with its 10-15 model parameters is a suitable model for this, but other simple bucket type models with a limited number of parameters could be used as well. We used the HBV model in the version HBV light (Seibert and Vis, 2012). The HBV model simulates streamflow based on time series of temperature and precipitation, as well as estimates of long-term evaporation using routines that represent snow accumulation and melt, soil moisture and groundwater storage and release, and streamflow routing. Model parameters for specific catchments are not measurable as they represent effective values at the catchment scale and are usually found by calibration or regionalization.

For the upper benchmark, we suggest the best possible streamflow simulation that can be achieved with a simple model. For this, the HBV model is calibrated using an optimization approach, in this case a genetic algorithm implemented in the HBV light software (Seibert, 2000). While better model performances might be possible with another model, differences in model performance among calibrated models are usually small (e.g., Refsgaard and Knudsen, 1996).

For the lower benchmark, we suggest two alternatives. In the fully uninformed case one might run the simple model with random parameters within feasible ranges. For the HBV model, such ranges have been suggested based on previous model applications in many catchments worldwide (Bergström, 1990; Lindström *et al.*, 1997; Seibert, 1999). Since single random parameter sets might result in varying model performances, an ensemble approach should be used. This means that a large number of random parameter sets is generated (here 1000 sets) and the model is run for each of them individually. The streamflow time series for the lower benchmark is then obtained by computing the mean streamflow from the 1000 simulated streamflow time series for each time step. If there are calibrated parameter sets for other (similar or nearby) catchments available, then these can be used as an alternative lower benchmark. Similar to the use of random parameter sets, no information from the catchment in question is used at all, but in this case the ensemble is compiled from calibrated parameter sets from other catchments. Again, the benchmark time series is computed as the ensemble mean.

Once streamflow time series for the upper and lower benchmarks have been established, the relative model performances of simulations for another model or parameterization can be determined. This relative performance measure, $R_{relative}$, allows evaluation of the model performance (R_x) relative to the performance of the upper and lower benchmarks (R_{upper} and R_{lower}) in a certain catchment and for a certain time period (Eq. 1).

$$R_{relative} = \frac{R_X - R_{lower}}{R_{upper} - R_{lower}} \quad (1)$$

If the model efficiency (Nash and Sutcliffe, 1970) is used as performance measure, $R_{relative}$ equals to the relative differences in the sums of squared errors S_ε (Eq. 2). Note also that this equation represents the Nash Sutcliffe efficiency exactly in case the mean observed discharge is used as the lower benchmark and the observed discharge as upper benchmark (i.e., S_ε =0).

$$R_{relative} = \frac{S_{\varepsilon,\,lower} - S_{\varepsilon,\,X}}{S_{\varepsilon,\,lower} - S_{\varepsilon,\,upper}} \quad (2)$$

Data sets

Two datasets were used to illustrate the variability in upper and lower benchmarks and therefore their usefulness in assessing model performance. One dataset consists of 673 catchments across the entire contiguous US (Newman *et al.*, 2014, 2015) and has been used in several modeling studies recently (Seibert and Vis, 2016; Guswa *et al.*, 2017; van Meerveld *et al.*, 2017). The other dataset includes 306 catchments in the UK (National River Flow Archive, 2014) and has also been used in several studies recently (Deckers *et al.*, 2010; Crooks *et al.*, 2014; Coxon *et al.*, 2015; Lewis *et al.*, 2018a, 2018b).

The HBV model was calibrated for each catchment individually to obtain the upper benchmark. For the lower benchmarks, mean streamflow time series were derived from two ensembles: 1) random parameter sets, and 2) regional parameter sets. The latter were the calibrated parameters for the other catchments in the dataset. To consider parameter uncertainty, we used 10 different optimized parameter sets from each of the other catchments in the US and UK datasets, respectively.

The benchmarks for the UK dataset were further used to evaluate the model performance of an uncalibrated SHETRAN model (Ewen *et al.*, 2000) with regard to streamflow. The SHETRAN model has been applied for 306 catchments in the UK without using any calibration or tuning against observed discharge and using only national datasets (including a national DEM and maps of hydrogeology, soils and land-cover) to derive parameter values (Lewis *et al.*, 2018a, 2018b). Potential evapotranspiration was calculated from the UK Climate Projections (UKCP09) 5km gridded climate variables (maximum and minimum daily temperature and monthly relative humidity, wind speed and sunshine hours data; Perry and Hollis, 2005) using the Penman-Monteith equation. Rainfall inputs were taken from the UKCP09 5km gridded daily rainfall dataset. Catchment boundary information and daily flow data were obtained from the UK National River Flow Archive (https://nrfa.ceh.ac.uk/). The simulations described here were conducted for 306 catchments for the period 1991-2002 (with a one-year spin-up period, 1990-1991). These catchments cover a broad range of the hydrological regimes in the UK and contain varying levels of human influence on streamflow. This model setup is known to perform poorly for groundwater dominated catchments as nationally available data do not appropriately parameterize or capture the heterogeneity of UK aquifers (Lewis *et al.*, 2018b).

Results

The benchmarks varied considerably for the different catchments in the US and the UK. For the upper benchmarks (i.e., when using calibrated parameters), model efficiency values were typically in the range

of 0.5 to 0.9 (Fig 1a and 2a, Table 1). The variability among the catchments was larger for the lower benchmarks (i.e., no catchment-specific calibration) with typical model efficiency values ranging from below 0 to above 0.8 (Fig 1b and 2c, Table 1). Some regional patterns can be observed for the upper and lower benchmarks, such as generally higher values of both benchmarks in the Pacific Northwest and along the Appalachian Mountains for the US catchments. For the UK catchments, the upper and lower benchmarks values were generally higher for catchments in western England, Wales and Scotland. However, despite these general trends, there was also considerable variability among neighboring catchments.

Comparison of the uncalibrated SHETRAN model simulations to the benchmarks showed that model performance was better on average than the lower benchmark using random parameters but somewhat poorer than the lower benchmark based on regional UK parameter sets (Fig 3, Table 1).

Discussion

The results clearly show that the lower and upper benchmarks differ among catchments and highlight why model simulations should not be compared to a fixed benchmark, such as the mean streamflow or a perfect model fit. As an example, the SHETRAN model performance for two UK catchments (Braan at Hermitage, 15023, and Ewe at Poolewe, 94001) is similar (0.85) but catchment 15023 has a lower benchmark (random) of 0.61 and an upper benchmark of 0.86, whereas catchment 94001 has a lower benchmark of 0.85 and an upper benchmark of 0.93. In other words, the model efficiency of 0.85 means an almost as good as possible performance for the first catchment, whereas the performance in the second catchment is just as good as what should be expected simply based on water balance constraints and would be achieved with a simple model without any calibration. Therefore, reporting only a model efficiency of 0.85 is insufficient for comparison of model performance across catchments.

There are obvious patterns in the spatial variations of the lower and upper benchmark (Fig 1 and 2), which are related to climate and geology. Higher precipitation and a larger proportion of snow, for instance, generally lead to a higher efficiency for the lower benchmarks. This is largely because for most wet catchments the actual evapotranspiration is relatively similar to the potential evapotranspiration and a model that adheres to the water balance will not produce results that are too dissimilar from the observations. On the other hand, for dry catchments the parameterization of the soil routine can lead to much more variable simulations with a largely varying split between evaporative fluxes and streamflow. Furthermore, in areas with deeper and more complex aquifers, such as in eastern and southern England, especially the lower benchmark values tend to be lower (Fig. 2). One reason is that streamflow series in these catchments are more damped and, thus, reflect precipitation time series less well. Data quality issues may also cause the values for the upper and lower benchmarks to vary. These include errors in the measurements, spatial variability in precipitation or the accuracy of the delineated catchment boundaries. Catchment complexity is another factor, but this is often compensated by the fact that larger catchments usually behave more linearly and are thus easier to model. Due to the combination of these different effects, the correlations of the values for the upper and lower benchmarks with single catchment characteristics were in general weak (Spearman rank coefficients were less than 0.4 for all correlations, except for the correlation between annual precipitation and the regional benchmark for the US data set (r_s=0.5)) and were not further analyzed.

Spatial patterns of relative model performance indicate where the SHETRAN model performs better/worse than could be expected. Theoretically, we would expect that the relative performance of the SHETRAN model would be higher in the groundwater dominated (Chalk) catchments in eastern England as SHETRAN has the capacity to represent the complex hydrogeology of the region. However, as discussed in Lewis et al. (2018a), the national datasets available could not adequately capture the heterogeneity of the model parameters for the Chalk and so the uncalibrated SHETRAN model performance was relatively poor. The comparison with the benchmarks therefore shows that the SHETRAN model did not lead to improved model simulations compared to a simple bucket-type model in terms of streamflow in these areas. One might question whether model simulations are useful in catchments where the performance is poorer than the lower benchmark. One argument to still use models in such cases is that whilst they might provide poorer streamflow simulations, they do so by better representing internal variables and fluxes. It also should be emphasized that there are of course reasons to use models beyond streamflow simulations, and for other variables the SHETRAN simulations still could be superior to simple models. For instance, simple models usually do not provide information on spatial variations in soil moisture, groundwater levels or streamflow.

For the lower benchmark, we suggested two different approaches. The difference between them is that in the first case parameter values are used that have been found acceptable in previous studies, whereas in the second case entire parameter sets are transferred and a parameter set is used if it performed well in at least one other catchment. The advantage of using random parameter values (within a certain range) for the lower benchmark is that this is a simple approach where no prior model calibrations are needed. This approach thus represents the fully uninformed case and is easiest to standardize for all catchments across the world (i.e. we can agree on which parameter ranges we use to determine the lower benchmark). The specified parameter ranges of course affect the values of this lower benchmark, but as an ensemble mean is used this effect is not as large as one may think. Initial tests suggest that even when the largest feasible ranges are used, model performances for the ensemble mean time series did not drop more than about 0.1 to 0.2 model efficiency units. The use of calibrated parameter sets from other catchments (i.e., the regional parameter sets) results in a more challenging lower benchmark and minimizes the effect of the initially chosen parameter value ranges. This approach is also closer to what one would do for streamflow simulations in an ungauged catchment but the actual efficiency of the lower benchmark may vary from region to region as it depends on the number of catchments for which these parameter sets are available (and their data quality). Of course, it would be unpractical to always first have to derive ensembles of parameter sets from other catchments before evaluating model performances in a certain catchment; one approach could be to agree on a large sample of parameter sets, i.e.,a global ensemble.

Conclusions

The examples from the US and UK clearly demonstrate a huge variation in terms of both how well streamflow can be simulated when a simple model is calibrated and how well the simple model performs with uncalibrated parameters. This clearly highlights the need for benchmarks as a complement for any goodmess-of-fit measure; without any comparison, it is not possible to fully judge how well a model performs. In the commonly used model efficiency (Nash and Sutcliffe, 1970), the mean observed streamflow is used as a lower benchmark, but this benchmark is not challenging enough to be really

informative. Similarly, the upper benchmark of a perfect fit is also not useful for comparing the results of different models because a perfect fit may not be possible due to data quality issues. Therefore, it would be useful for the hydrological modeling community to agree on benchmarks to be used for the comparison of model performance. We suggest using a simple bucket-type model with few model parameters to determine both the lower and the upper benchmarks and argue that the use of an upper and lower benchmark is a valuable way to assess model performance.

It might be difficult to agree on specific benchmarks as there are many choices, such as which simple model to use or how to define the lower benchmark. However, we argue that regardless of the specific choices, the use of lower and upper benchmarks defined in some reasonable way is better than using simplistic rules of a thumb, such as a model performance with a model efficiency value above 0.7 being good without any comparison to what is possible and can be expected. As illustrated by the evaluation of the SHETRAN model simulations for a large number of catchments in the UK, the use of benchmarks is needed to assess model performance. We encourage the hydrological modeling community to further discuss and explore the use of lower and upper benchmarks. Hopefully this will lead to generally accepted benchmarks in the future, similar to the common use of the model efficiency in the past decades.

References

Bergström S. 1976. Development and application of a conceptual runoff model for Scandinavian catchments. SMHI, Norrköping, Sweden, No. RHO 7, 134 pp.

Bergström S. 1990. Parametervärden för HBV-modellen i Sverige, Erfarenheter från modelkalibreringar under perioden 1975- 1989 (in Swedish, English title: (Parameter values for the HBV model in Sweden, experiences from model calibrations 1975-1989). SMHI, Norrköping, Sweden, Hydrologi No 28, 35 pp.

Coxon G, Freer J, Westerberg IK, Wagener T, Woods R, Smith PJ. 2015. A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations. *Water Resources Research* **51**: 5531–5546 DOI: 10.1002/2014WR016532

Crooks SM, Kay AL, Davies HN, Bell VA. 2014. From Catchment to National Scale Rainfall-Runoff Modelling: Demonstration of a Hydrological Modelling Framework: 63–88 DOI: 10.3390/hydrology1010063

Deckers DLEH, Booij MJ, Rientjes THM, Krol MS. 2010. Catchment Variability and Parameter Estimation in Multi-Objective Regionalisation of a Rainfall – Runoff Model. *Water Resources Management* **24**: 3961–3985 DOI: 10.1007/s11269-010-9642-8

Ewen J, Parkin G, O'Connell PE. 2000. SHETRAN : Distributed river basin flow and transport modeling system. *Journal of Hydrologic Engineering* **5** (JULY): 250–258

Garrick M, Cunnane C, Nash JE. 1978. A criterion of efficiency for rainfall-runoff models. *Journal of Hydrology* **36**: 375–381

Guswa AJ, Hamel P, P.James D-F. 2017. Potential effects of landscape change on water supplies in the presence of reservoir storage. *Water Resources Research* **53**: 2679–2692 DOI: 10.1002/2016WR019691

Lewis E, Birkinshaw S, Kilsby C, Fowler H. 2018a. A physically-based hydrological modelling system for Great Britain. *Environmental Modelling and Software*: in review

Lewis E, Birkinshaw S, Kilsby C, Fowler HJ. 2018b. Collective assessment of a new physically based hydrological modelling system for national scale applications. *Water Resources Research*: in review

Lindström G, Johansson B, Persson M, Gardelin M, Bergström S. 1997. Development and test of the distributed HBV-96 hydrological model. *Journal of Hydrology* **201**: 272–288

van Meerveld HJ, Vis MJP, Seibert J. 2017. Information content of stream level class data for hydrological model calibration. *Hydrology and Earth System Sciences* **21** (9): 4895–4905 DOI: 10.5194/hess-21-4895-2017

Moriasi DN, Arnold JG, Van Liew MW, Bingner RL, Harmel RD, Veith TL. 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE* **50** (3): 885–900 DOI: 10.13031/2013.23153

Nash JE, Sutcliffe J V. 1970. River flow forecasting through conceptual models Part I- A discussion of principles. *Journal of Hydrology* **10**: 282–290 DOI: 10.1016/0022-1694(70)90255-6

National River Flow Archive. 2014. Hydrometric areas for Great Britain and Northern Ireland. NERC-Environmental Information Data Centre DOI: 10.5285/1957166d-7523-44f4-b279-aa5314163237

Newman AJ, Clark MP, Sampson K, Wood A, Hay LE, Bock A, Viger RJ, Blodgett D, Brekke L, Arnold JR, et al. 2015. Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences* **19** (1): 209–223 DOI: 10.5194/hess-19-209-2015

Newman AJ, Sampson K, Clark MP, Bock A, Viger RJ, Blodgett D. 2014. A large-sample watershed-scale hydrometeorological dataset for the contiguous USA. Boulder, CO: UCAR/NCAR Available at: http://dx.doi.org/10.5065/D6MW2F4D

Pappenberger F, Ramos M, Cloke HL, Fredrik W. 2014. How do I know if my forecasts are better ? Using benchmarks in Hydrological Ensemble Predictions. *JOURNAL OF HYDROLOGY* **16**: 12251 DOI: 10.1016/j.jhydrol.2015.01.024

Perry M, Hollis D. 2005. The generation of monthly gridded datasets for a range of climatic variables over the UK. *International Journal of Climatology* **25**: 1041–1054 DOI: 10.1002/joc.1161

Refsgaard JC, Knudsen J. 1996. Operational Validation and Intercomparison of Different Types of Hydrological Models. *Water Resources Research* **32** (7): 2189–2202 DOI: 10.1029/96WR00896

Ritter A, Muñoz-Carpena R. 2013. Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments. *Journal of Hydrology* **480**: 33–45 DOI: 10.1016/j.jhydrol.2012.12.004

Schaefli B, Gupta H V. 2007. Do Nash values have value ? *Hydrological Processes* **21**: 2075–2080 DOI: 10.1002/hyp

Seibert J. 1999. Regionalisation of parameters for a conceptual rainfall-runoff model. *Agricultural and Forest Meteorology* **98–99**: 279–293

Seibert J. 2000. Multi-criteria calibration of a conceptual runoff model using a genetic algorithm. *Hydrology and Earth System Sciences* **4** (2): 215–224 DOI: 10.5194/hess-4-215-2000

Seibert J. 2001. On the need for benchmarks in hydrological modelling. *Hydrological Processes* **15** (6): 1063–1064 DOI: 10.1002/hyp.446

Seibert J, Vis MJP. 2012. Teaching hydrological modeling with a user-friendly catchment-runoff-model software package. *Hydrology and Earth System Sciences* **16** (9): 3315–3325 DOI: 10.5194/hess-16-3315-2012

Seibert J, Vis MJP. 2016. How informative are stream level observations in different geographic regions? *Hydrological Processes* **30** (14): 2498–2508 DOI: 10.1002/hyp.10887

Tables

Table 1. Ranges of model efficiency values between the 10th and 90th percentile as well as the median (in parentheses) for the catchments in the US and UK datasets for the different benchmarks and the uncalibrated SHETRAN model

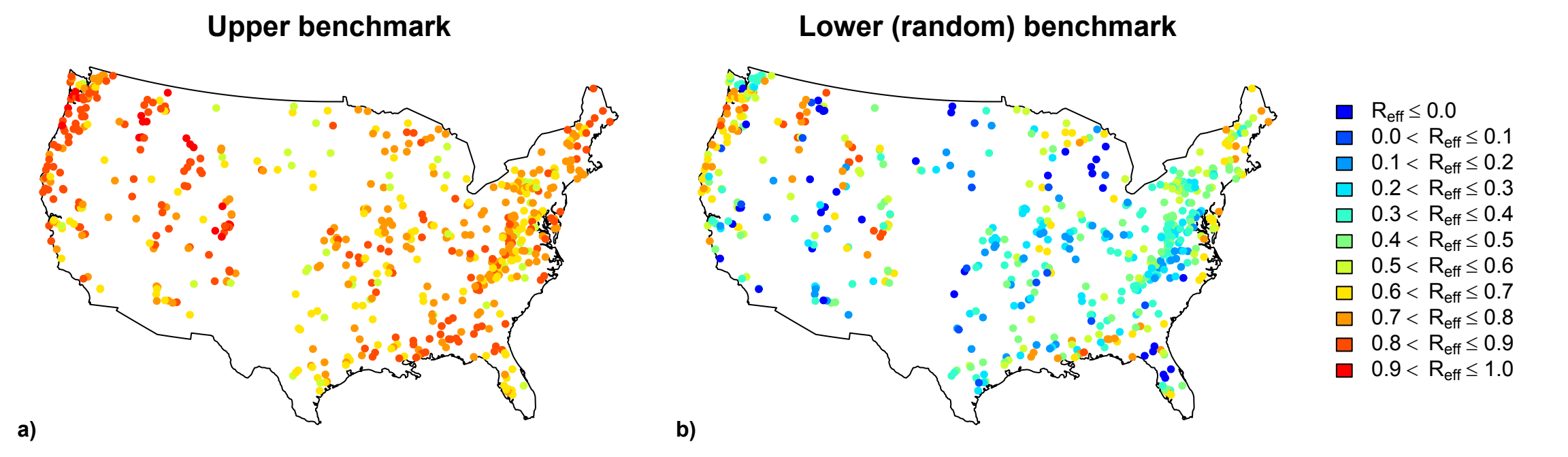
	US dataset	UK dataset
Upper benchmark (calibrated parameter values)	0.52 to 0.87 (0.74)	0.75 to 0.92 (0.85)
Lower benchmark (regional parameter values)	-0.27 to 0.77 (0.54)	-0.68 to 0.85 (0.72)
Lower benchmark (random parameter values)	-0.06 to 0.70 (0.38)	0.13 to 0.81 (0.52)
Uncalibrated SHETRAN model	-	-2.26 to 0.84 (0.67)

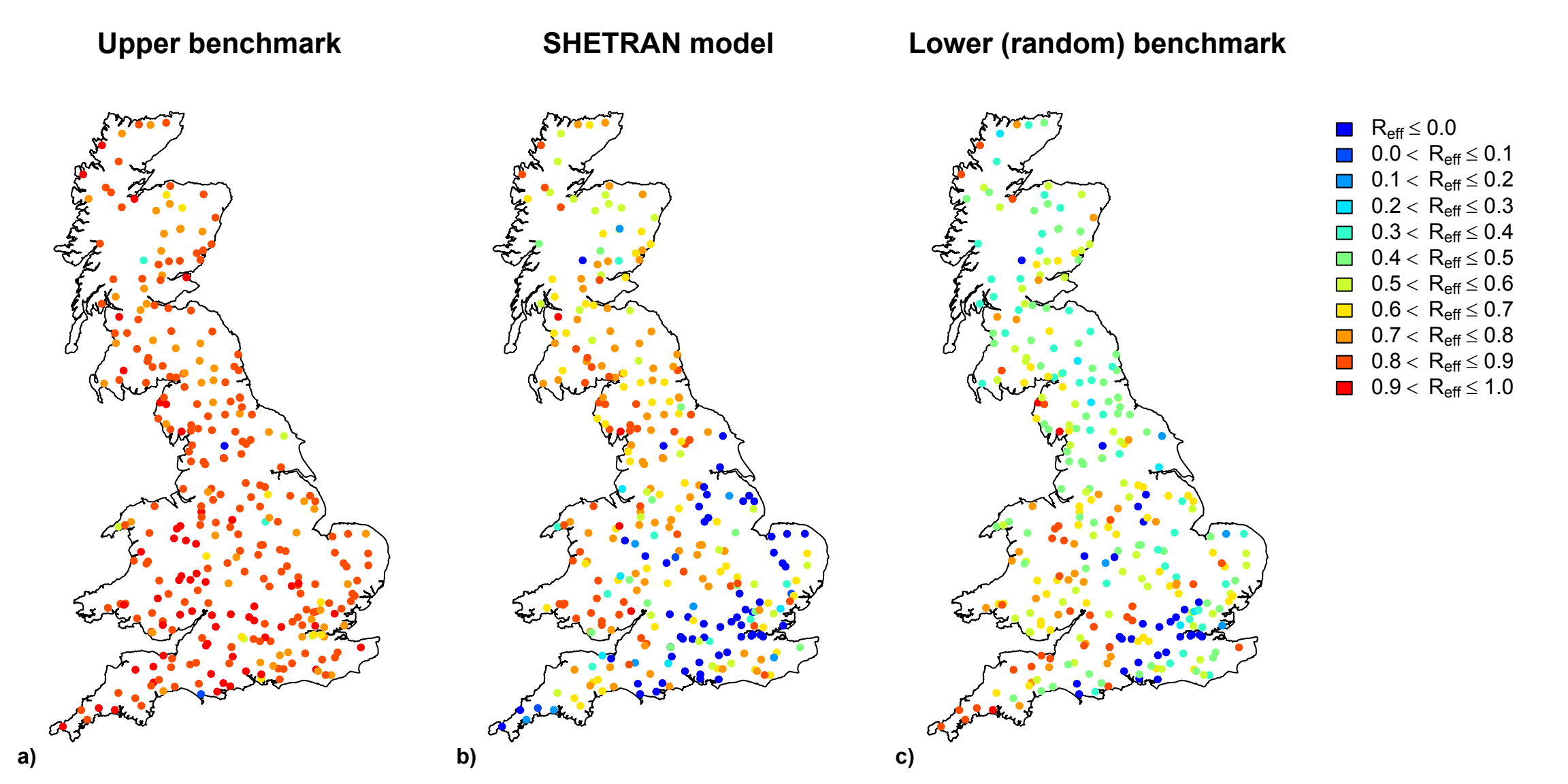
Figures:

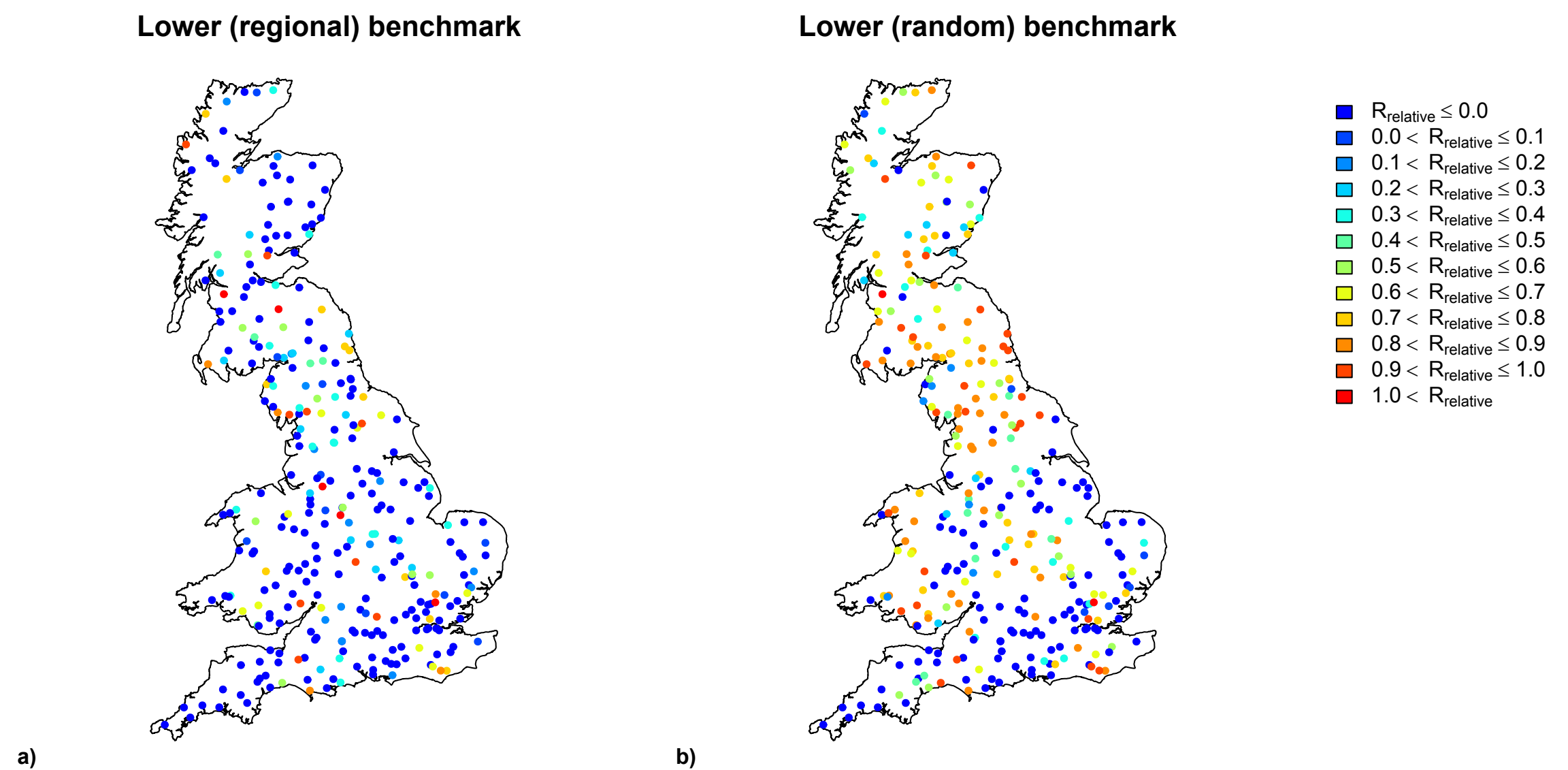
Fig 1: Upper (a, left) and lower (b, right) benchmarks for US catchments. The latter was based on random parameters. Patterns were similar when regional parameters were used to obtain the lower benchmark but model efficiency values were generally higher (Table 1).

Fig 2: Upper (a, left) and lower (c, right) benchmarks for UK catchments, as well as performance of the uncalibrated SHETRAN model (b, middle). The lower benchmark was based on random parameters. Patterns were similar when regional parameters were used to obtain the lower benchmark but model efficiency values were generally higher (Table 1).

Fig.3: Relative model performances of the SHETRAN model. For the lower benchmark two different ensembles were used, which were based on regional (a, left) and random (b, right) parameter sets.


Upper benchmark
Lower (random) benchmark
a)
b)
$R_{eff} \leq 0.0$
$0.0 < R_{eff} \leq 0.1$
$0.1 < R_{eff} \leq 0.2$
$0.2 < R_{eff} \leq 0.3$
$0.3 < R_{eff} \leq 0.4$
$0.4 < R_{eff} \leq 0.5$
$0.5 < R_{eff} \leq 0.6$
$0.6 < R_{eff} \leq 0.7$
$0.7 < R_{eff} \leq 0.8$
$0.8 < R_{eff} \leq 0.9$
$0.9 < R_{eff} \leq 1.0$


Upper benchmark
SHETRAN model
Lower (random) benchmark
$R_{eff} \leq 0.0$
$0.0 < R_{eff} \leq 0.1$
$0.1 < R_{eff} \leq 0.2$
$0.2 < R_{eff} \leq 0.3$
$0.3 < R_{eff} \leq 0.4$
$0.4 < R_{eff} \leq 0.5$
$0.5 < R_{eff} \leq 0.6$
$0.6 < R_{eff} \leq 0.7$
$0.7 < R_{eff} \leq 0.8$
$0.8 < R_{eff} \leq 0.9$
$0.9 < R_{eff} \leq 1.0$
a)
b)
c)


Lower (regional) benchmark
Lower (random) benchmark
$R_{relative} \leq 0.0$
$0.0 < R_{relative} \leq 0.1$
$0.1 < R_{relative} \leq 0.2$
$0.2 < R_{relative} \leq 0.3$
$0.3 < R_{relative} \leq 0.4$
$0.4 < R_{relative} \leq 0.5$
$0.5 < R_{relative} \leq 0.6$
$0.6 < R_{relative} \leq 0.7$
$0.7 < R_{relative} \leq 0.8$
$0.8 < R_{relative} \leq 0.9$
$0.9 < R_{relative} \leq 1.0$
$1.0 < R_{relative}$
a)
b)