

**Reference:** Gragne, A. S., Sharma, A., Mehrotra, R., and Alfredsen, K.: Improving inflow forecasting into hydropower reservoirs through a complementary modelling framework, Hydrol. Earth Syst. Sci. Discuss., 11, 12063-12101, doi:10.5194/hessd-11-12063-2014, 2014.

Dear Editor - dear Dr. Toth,

First of all, we would like to thank you for favourably receiving our manuscript for publication subject to minor revision.

We also would like to thank the two anonymous reviewers for the constructive comments and suggestions they provided in this round of review as well. We believe the comments and suggestions have helped improve the quality of the manuscript.

We have revised the manuscript in accordance with the comments and suggestions of the reviewers and the three comments forwarded by the editor (Dr. Toth). As per the suggestions, we have further evaluated the forecasting skill by plotting PIT uniform plots, estimated the transformation parameters (and AR model coefficients) by means of a likelihood approach, and addressed other issues raised by reviewers. We have added a subsection just before the conclusions wherein we briefly compared results of the present approach with the likelihood based approach.

In the following we present the comments from the reviewers, our responses to the issues raised, and the changes made to the manuscript. The additional literatures we cited are listed at the end.

## Referee #1

### Comments:

1. *It is suggested the title of paper is changed as “Improving real-time inflow forecasting for hydropower reservoir”.*

We thank referee #1 for the suggestion. We have changed the title.

Revised Manuscript, (Title), Page 1 (Line 1):

Improving real-time inflow forecasting into hydropower reservoirs through a complementary modelling framework

2. *Only CR is used to evaluate the prediction results. More indexes should be added to identify the overall quality of prediction intervals.*

We thank referee #1 for the suggestion. In the revised manuscript we have used the probability integral transform (PIT) plot (also known as predictive QQ plot) as an additional probabilistic forecast verification approach.

Revised Manuscript, Section 1, Page 5 (Line 8-12):

In this paper, the improvement levels achieved are evaluated deterministically using the same or similar metrics as past studies, and probabilistically using: (i) the containing ratio (Xiong et al., 2009), which is also referred to as reliability score (e.g. Renard et al., 2010); and (ii) the probability integral transform (PIT) plot, which Thyer et al. (2009) refer to as the predictive QQ plot.

According to Xiong et al. (2009), of all indices the containing ratio (CR), which describes the percentage of observed inflows falling in the desired interval percentage, is the widely used metrics for assessing reliability of probabilistic forecasts. We adopt the CR metric for describing the reliability of the forecasts with the desired interval percentage of 95% ( $\alpha = 0.05$ ). Beside the CR, we verify the probabilistic forecasts graphically using the less formal PIT uniform probability plot. The working procedure along detailed application examples can be found in Laio and Tamea (2007) and Thyer et al. (2009). Among others, Pokhrel et al. (2013) and Wang et al. (2009) demonstrate viability of the 'PIT uniform probability plot' approach for checking uniformity (and investigating the causes, in cases of deviations from uniformity) without binning the data subjectively.

It can also be noted that for shorter forecast lead-times, the percentage of observations contained in the prediction bounds exceed 95%. Although a greater proportion of observations falling in the prediction bound is desirable, a high CR at short forecast lead-times might indicate a too wide band-width. This along a CR that declines with increased lead-time might suggest invalidity of the assumptions behind computation of the bounds (e.g. Smith et al., 2012). The two issues at stake here are the Gaussian assumption on the basis of which the prediction bounds were constructed, and the model identification and parameter estimation approach implemented. However, the results from the PIT uniformity probability test (Fig. 9c-f) confirm that the distribution of the forecasts from the complementary framework are consistent with the observed inflow. The PIT plots of the observed inflows almost plot along the diagonal bisector and remain within the Kolmogorov 5% significance in cases of slightly deviations. From operational hydrology point of view, we concur with the opinion of Thyer et al. (2009) that the toughest goodness-of-fit test the complementary framework has to pass is whether the predictive distribution is consistent with the observed inflow, which the PIT uniform probability plots evaluate directly. With regard to the model assumptions, we recall from the Kolmogorov-Smirnov test (section 3.3) that the residuals from the error forecasting model did not honour the homoscedasticity assumption even after transformation, which might suggest invalidity of the assumption. According to Schoups and Vrugt (2010), in hydrologic applications residual series are often assumed to be independent and identically distributed but these assumptions are usually violated. In the next section, we briefly assess reliability of the model identification and parameter estimation approach implemented in this study.

3. *P16, Line 18 [marked up manuscript]. Why the NSE value in calibration period is smaller than that in the validation period?*

We thank the reviewer for the question. The *NSE* value associated to calibration of the error model is smaller than the *NSE* for the validation period. It has to also be noted that, the efficiency of the HBV model during the calibration period is higher than the efficiency in the validation period. We believe that the former might imply extraction of as much information as possible from the data by the conceptual model during the model calibration. Another possible explanation could be discrepancy in the hydrologic condition of the study are in the calibration and validation period. We have attempted to give the same remark on the revised manuscript.

Calibration efficiencies calculated for the error model using the RMSE, RE and NSE metrics are 0.096, -100% and 0.517, respectively. Corresponding values for the validation period are computed as 0.095, 20.3% and 0.630, respectively. NSE values for the calibration and validation periods imply ability of the error model to capture at least half of the discrepancies observed between observations and predictions from the conceptual model. All the three metrics reveal a higher efficiency in the validation set than the calibration set. With reference to Table 2, this suggests too much fitting of the HBV model to the data that led to extraction of more information from the calibration set. Assessment of the residuals from the complementary framework reveals that the transformation reduced the maximum deviation between the theoretical and the sample lines slightly from 0.13 to 0.10; yet the residuals are not normally distributed (i.e. Kolmogorov-Smirnov statistic of 0.008 at significance level of  $\alpha = 0.05$ ).

4. *Explain why the accuracy of the predicted inflow is decreased (or increased) with the increase of lead time in Table 5?*

We thank the reviewer for pointing out this issue. The results from Table 5 reveal that the CR of the prediction interval is very high at the short forecast lead-times and declines at longer lead-times. One possible explanation could be invalidity of the assumption made for estimating the prediction interval (e.g. Smith et al., 2012). We have attempted to make the same remark in revised manuscript.

Revised Manuscript, Section 3, Page 19 (Line 9-19):

Similarly, detailed assessment of the reliability (Table 5) shows that the CR of the forecasting system can get below 95% at forecast lead-times less than 17 hours; e.g. at forecast lead-time of 9 hours only 89% of the observed spring inflows of year 2006/07 are bracketed in the 95% prediction interval. It can also be noted that for shorter forecast lead-times, the percentage of observations contained in the prediction bounds exceed 95%. Although a greater proportion of observations falling in the prediction bound is desirable, a high CR at short forecast lead-times might indicate a too wide band-width. This along a CR that declines with increased lead-time might suggest invalidity of the assumptions behind computation of the bounds (e.g. Smith et al., 2012). The two issues at stake here are the Gaussian assumption on the basis of which the prediction bounds were constructed, and the model identification and parameter estimation approach implemented.

5. *The quality of figures should be improved.*

We thank reviewer #1 for the suggestion. We have reproduced the figures with 600 dpi resolution.

## Referee #2

### Comments:

1. *(Line 1 Page 30) Table 5 suggests that reliability is an issue of the forecasts issued from this study. The CR should be close to 95% to reflect that the residual distribution honours the assumptions. However, the CR is 99.9% at a 1-hr lead-time for all cases, suggesting the confidence intervals are too wide. To investigate the reliability issue, you can compare the observed quantile of  $\eta_t$  (Line 16 Page 8) with the corresponding quantile fitted by the normal distribution. Alternatively, you examine the PIT uniformity plot. A reliable forecast is the first importance at an hourly step, because the single-value forecasts based on the time persistence can be very accurate.*

We agree with referee #2 and thank the reviewer for the suggestion. We have used the PIT uniform probability plot in the revised manuscript.

Revised Manuscript, Section 1, Page 5 (Line 8-12):

In this paper, the improvement levels achieved are evaluated deterministically using the same or similar metrics as past studies, and probabilistically using: (i) the containing ratio (Xiong et al., 2009), which is also referred to as reliability score (e.g. Renard et al., 2010); and (ii) the probability integral transform (PIT) plot, which Thyer et al. (2009) refer to as the predictive QQ plot.

Revised Manuscript, Section 2, Page 10 (Line 12-21):

According to Xiong et al. (2009), of all indices the containing ratio (CR), which describes the percentage of observed inflows falling in the desired interval percentage, is the widely used metrics for assessing reliability of probabilistic forecasts. We adopt the CR metric for describing the reliability of the forecasts with the desired interval percentage of 95% ( $\alpha = 0.05$ ). Beside the CR, we verify the probabilistic forecasts graphically using the less formal PIT uniform probability plot. The working procedure along detailed application examples can be found in Laio and Tamea (2007) and Thyer et al. (2009). Among others, Pokhrel et al. (2013) and Wang et al. (2009) demonstrate viability of the 'PIT uniform probability plot' approach for checking uniformity (and investigating the causes, in cases of deviations from uniformity) without binning the data subjectively.

It can also be noted that for shorter forecast lead-times, the percentage of observations contained in the prediction bounds exceed 95%. Although a greater proportion of observations falling in the prediction bound is desirable, a high CR at short forecast lead-times might indicate a too wide band-width. This along a CR that declines with increased lead-time might suggest invalidity of the assumptions behind computation of the bounds (e.g. Smith et al., 2012). The two issues at stake here are the Gaussian assumption on the basis of which the prediction bounds were constructed, and the model identification and parameter estimation approach implemented. However, the results from the PIT uniformity probability test (Fig. 9c-f) confirm that the distribution of the forecasts from the complementary framework are consistent with the observed inflow. The PIT plots of the observed inflows almost plot along the diagonal bisector and remain within the Kolmogorov 5% significance in cases of slightly deviations. From operational hydrology point of view, we concur with the opinion of Thyer et al. (2009) that the toughest goodness-of-fit test the complementary framework has to pass is whether the predictive distribution is consistent with the observed inflow, which the PIT uniform probability plots evaluate directly. With regard to the model assumptions, we recall from the Kolmogorov-Smirnov test (section 3.3) that the residuals from the error forecasting model did not honour the homoscedasticity assumption even after transformation, which might suggest invalidity of the assumption. According to Schoups and Vrugt (2010), in hydrologic applications residual series are often assumed to be independent and identically distributed but these assumptions are usually violated. In the next section, we briefly assess reliability of the model identification and parameter estimation approach implemented in this study.

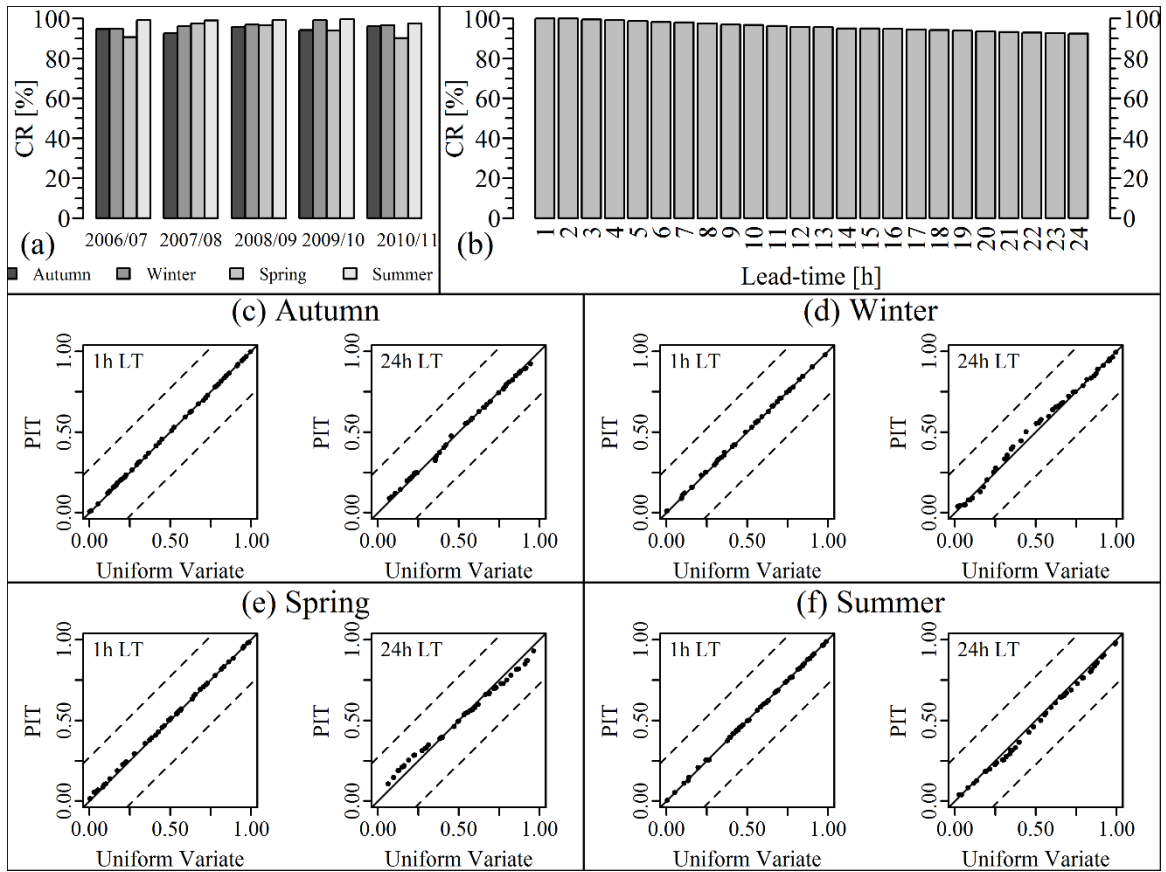


Figure 1. Reliability score (containing ratio-CR) for 95% prediction interval for: (a) each season of every hydrologic year; and (b) different forecast lead-times based on entire series. Panels (c)-(f): sample PIT uniform probability plots for each of the four seasons at 1 hour and 24 hour forecast lead-times. Solid line designates the theoretical uniform distribution, broken lines represent the Kolmogorov significance band, and the dots denote PIT value of the observed inflow.

2. (Lines 18-21 Page 8) *I guess that the reliability issue is related to the estimation of the transformation parameters used in this study. The estimation is somewhat based on the visual inspection and can be definitely improved by a likelihood approach. You can easily form the likelihood as a function of  $a_1, \dots, a_p$  and two transformation parameters and maximise the log-likelihood function to estimate all parameters. The likelihood approach ensures the distributional assumption is statistically used in the estimation. I hope the likelihood estimation can solve the reliability issue mentioned above.*

We thank referee #2 for the suggestion. We have re-estimated the transformation parameters and coefficients of the AR model using a likelihood approach and compared the result with that of the present approach. Deterministic metrics (Table A1) show that the likelihood based approach is superior to the present approach. We afraid that this has not led to improved characterization of the uncertainty (please refer to Figures A1-A11). But we are sure that it attests that the heteroscedasticity and disproportional percentage of observations contained in the 95% prediction interval did not uniquely associate to the calibration technique employed.

Revised Manuscript, Section 3, Page 20 (Line 1-32) - Page 21 (Line 1-3):

### **3.6 On the implemented parameter estimation technique**

The parameter (AR model coefficient(s) and transformation parameters) estimation technique we employed (section 2.2.2) follows a pseudo multi-objective optimization approach, which includes minimizing the sum of squares of the residuals and making sure a homoscedastic residual series. We first employed the Least Square (LS) method to estimate the parameters associated to several AR models (of orders 1 to 3). Since the unit of the inflows (the errors as well) in the transformed space depended on the transformation parameters, and the inclusion of the transformation parameters into the calibration problem posed a challenge to identify the optimal among the candidate AR models, we resorted to the dimensionless Kolmogorov-Smirnov (KS) statistic. The KS metric served as a relative quantitative measure to discriminate between candidate models by measuring how close-to-constant the residual variances' are. As a result, the selected AR model is suboptimal in terms of yielding the least discordance between predictions and observations. Putting aside the issue of (in)validity of the Gaussian assumption, we demonstrate that shortcomings of the present LS and KS (LS-KS) model the probabilistic metrics revealed are not unique to the implemented parameter estimation approach. In order to verify this, we setup an AR model estimated the coefficients and transformation parameters by maximizing the Gaussian maximum likelihood (GML).

An AR(2) model was identified with coefficients and transformation parameters:  $\beta = 1.08$ ,  $\lambda = 0.01$ ,  $a_1 = 1.82$  and  $a_2 = -0.82$ . All the deterministic metrics used in this study confirm performance improvement of a slight degree by the GML based model during calibration and validation. This does not come as a surprise because parameters of the LS-LK based model were suboptimal. On the other hand, the KS test revealed that the maximum distance between the sample line and the theoretical line increased to 0.290, which is higher than the statistic the error transformation using parameterization of the LS-KS model (0.10) yielded. To be fair, comparison of the KS statistics associated to the GML and LS-LK transformation parameters might not be appropriate because the LS-KS based AR model was selected for its low KS statistic. Nevertheless, the KS statistic corresponding to the GML based transformation shows a heteroscedasticity of degree higher than the untransformed residuals (0.13). The PIT uniform probability plots assured that the distribution of the predictions matches the observation very well. Comparison of the CR of the GML and LS-LK based models showed a similar proportion of observations contained in the prediction interval. The CR again reveals the same characteristics of high values at short lead-times and the fraction of observations contained in the prediction bound declines at longer lead-times. This affirms that validity of the Gaussian assumptions stand out as the main issue requiring further investigation in relation to probabilistic assessment.

Table A1. Tabular comparison of performance of the present (LS-KS) and the GML based forecasting system over the validation period.

	<i>RMSE</i>		<i>RE</i>		<i>NSE</i>	
	LS-KS	GML	LS-KS	GML	LS-KS	GML
Overall	0.095	0.085	3.8	4.4	0.896	0.908
Autumn	0.094	0.079	2.9	3	0.904	0.926
Winter	0.090	0.08	0.7	-0.4	0.905	0.915
Spring	0.132	0.121	4.6	6.5	0.859	0.873
Summer	0.044	0.036	10.2	10.7	0.873	0.892

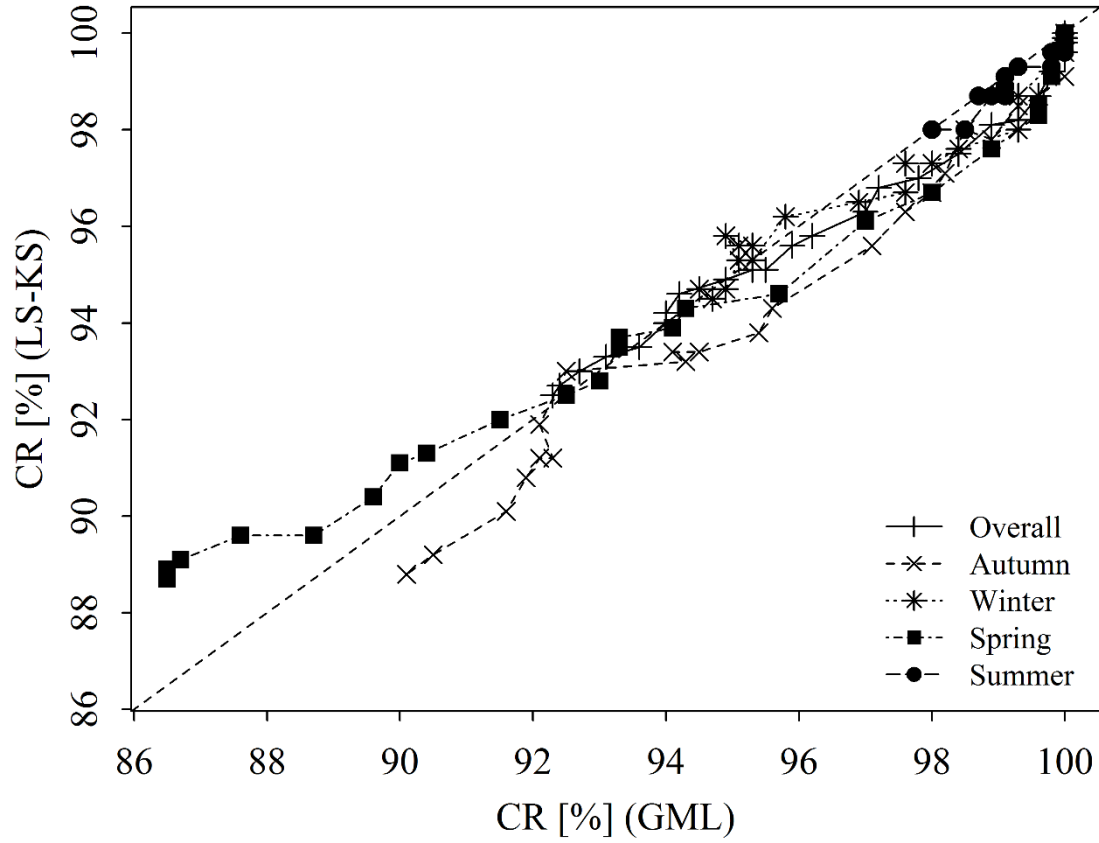


Figure A1. Graphical comparison of the CR of the GML and LS-KS (present) for different seasons (NB: each symbols designates a single forecast lead-time step).

We would be happy to answer any further question!

Best regards,

Ashenafi S Gagne (corresponding author)

## References

- Laio, F., and Tamea, S.: Verification tools for probabilistic forecasts of continuous hydrological variables, *Hydrol. Earth Syst. Sci.*, 11(4), 1267 – 1277, 2007.
- Pokhrel, P., Robertson, D.E. and Wang, Q.J.: A Bayesian joint probability post-processor for reducing errors and quantifying uncertainty in monthly streamflow predictions. *Hydrol. Earth Syst. Sci. Discuss.* 9(10), 11199-11225, 2012.
- Schoups, G. and Vrugt, J.A.: A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. *Water Resources Research* 46(10), 2010.
- Thyer, M., Renard, B., Kavetski, D., Kuczera, G., Franks, S. W., and Srikanthan, S.: Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: a case study using Bayesian total error analysis, *Water Resour. Res.*, 45, W00B14, doi:10.1029/2008WR006825, 2009.
- Wang, Q. J., Robertson, D. E., and Chiew, F. H. S.: A Bayesian joint probability modeling approach for seasonal forecasting of streamflows at multiple sites, *Water Resour. Res.*, 45, W05407, doi:10.1029/2008WR007355, 2009.

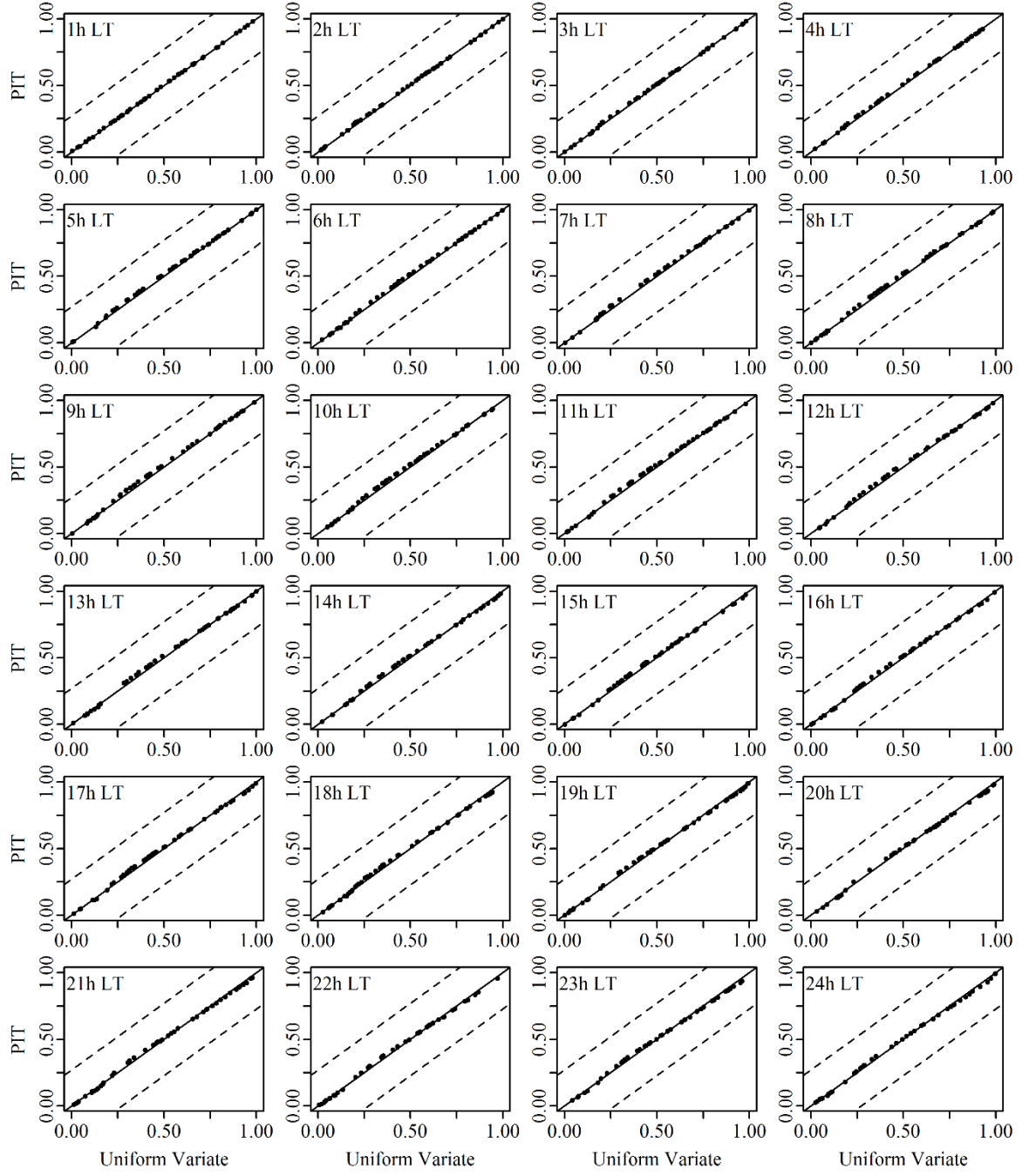


Figure A2. PIT uniform probability plot (entire data set; LS-KS (present) approach).



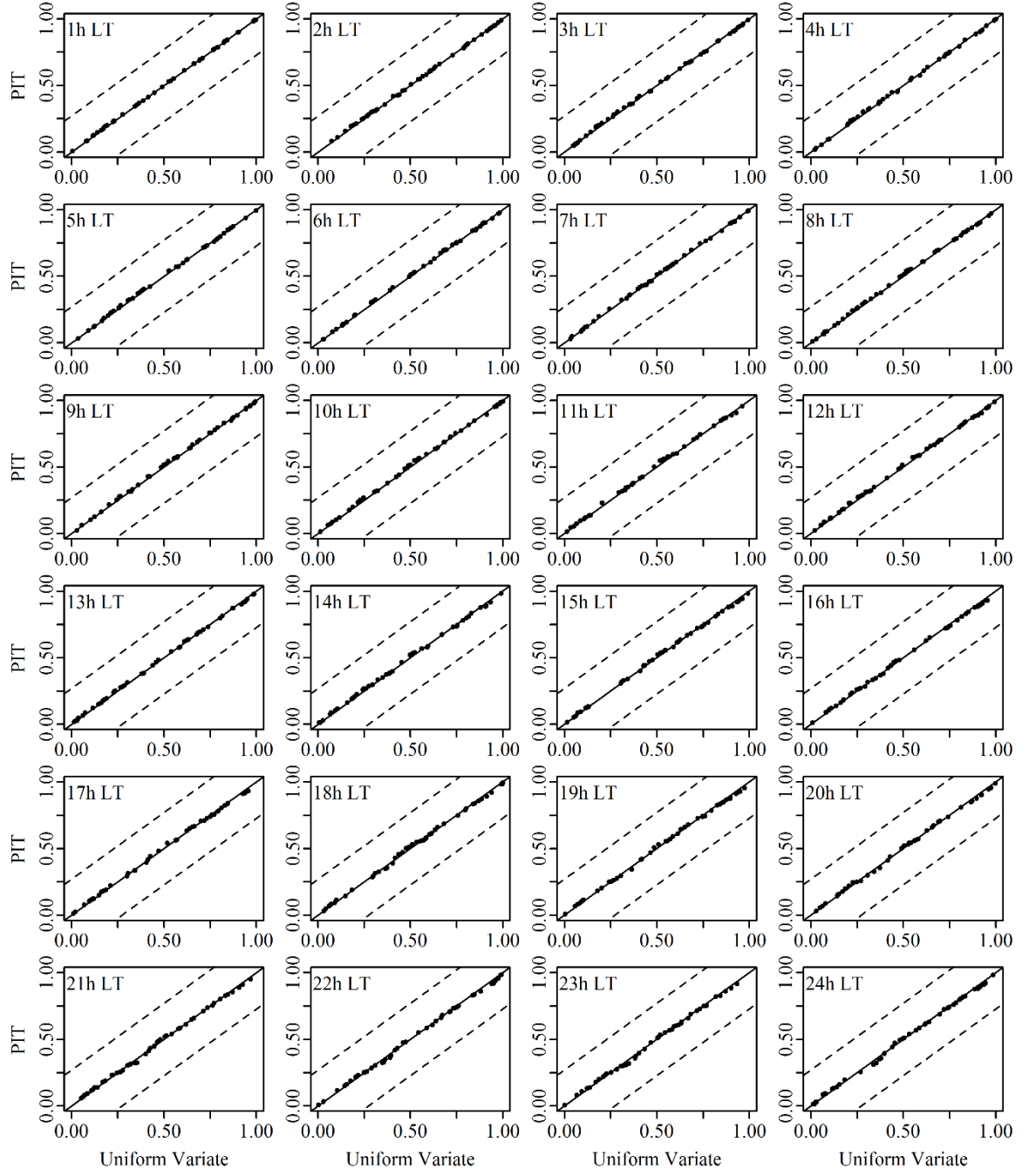


Figure A3. PIT uniform probability plot (autumn set; LS-KS (present) approach).

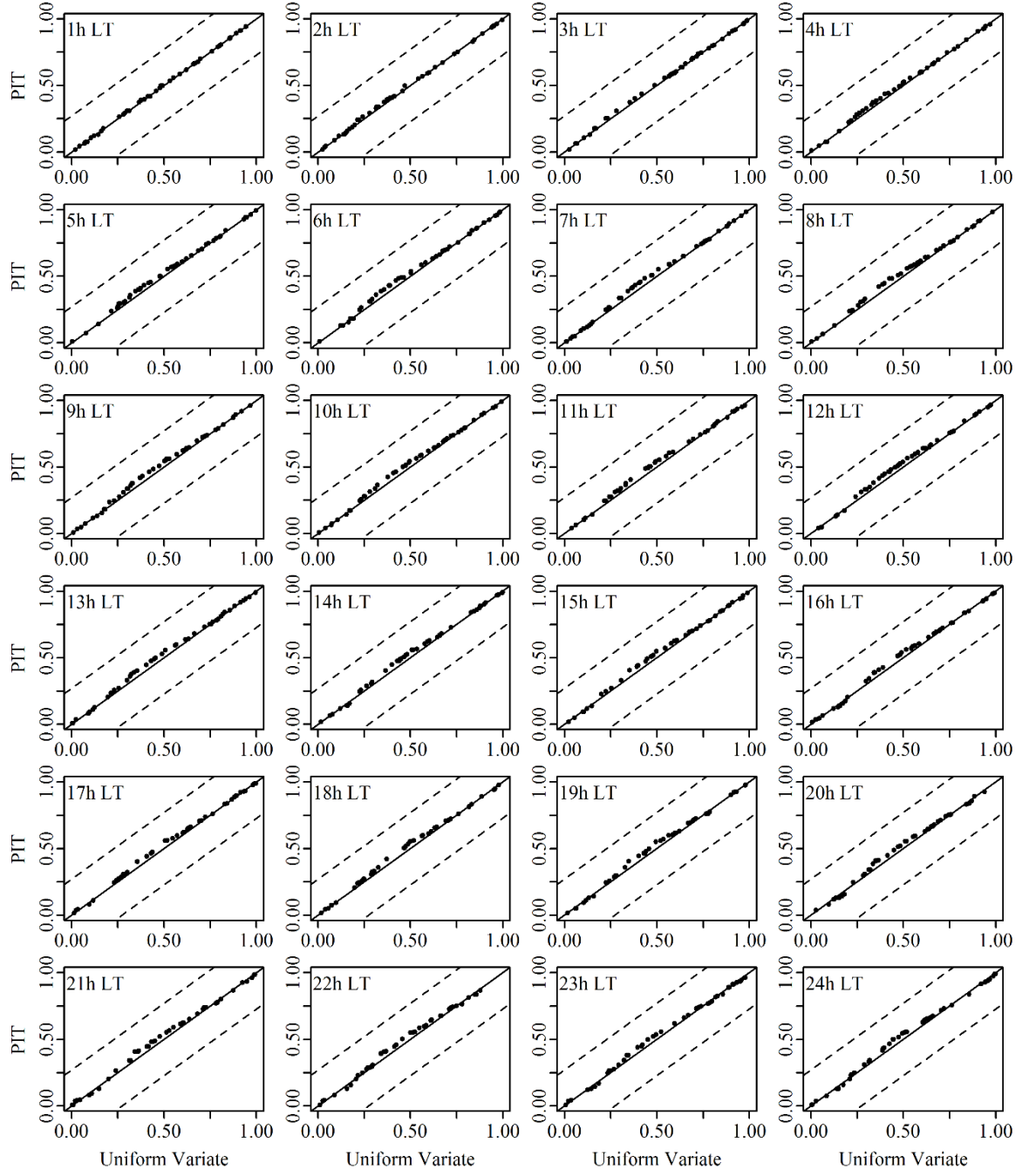


Figure A4. PIT uniform probability plot (winter set; LS-KS (present) approach).

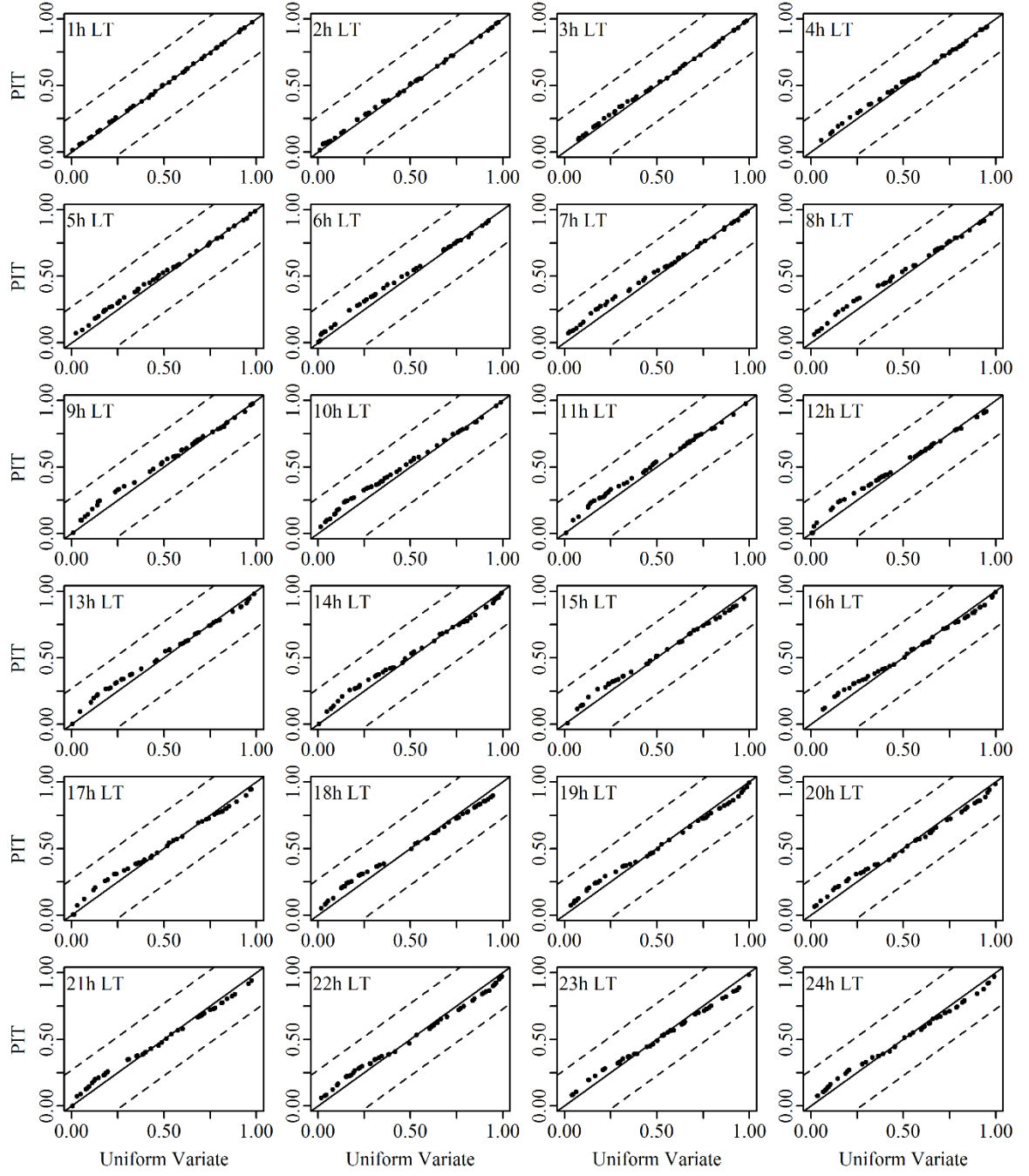


Figure A5. PIT uniform probability plot (spring set; LS-KS (present) approach).

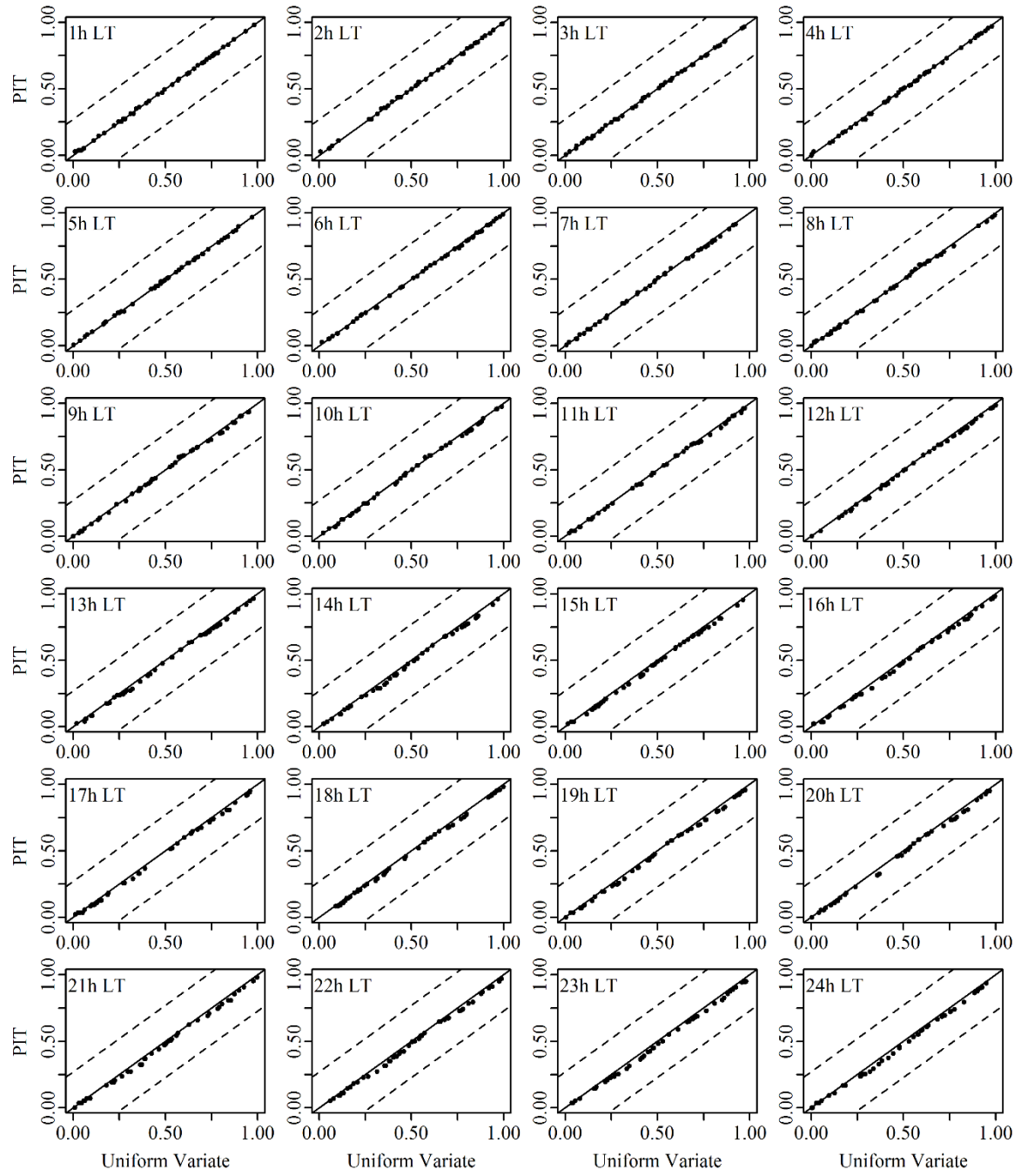


Figure A6. PIT uniform probability plot (summer set; LS-KS (present) approach).

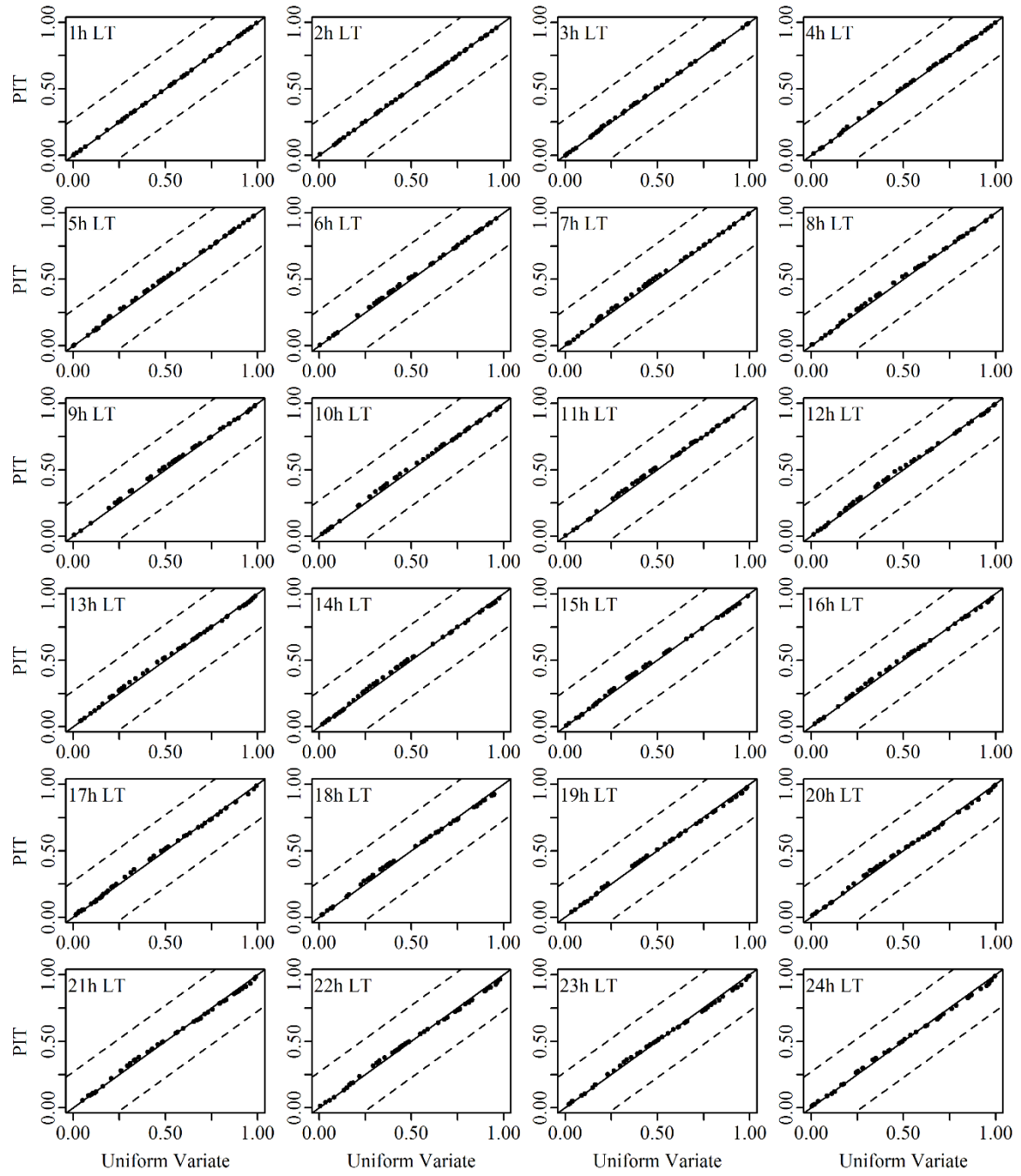


Figure A7. PIT uniform probability plot (entire data set; GML approach).

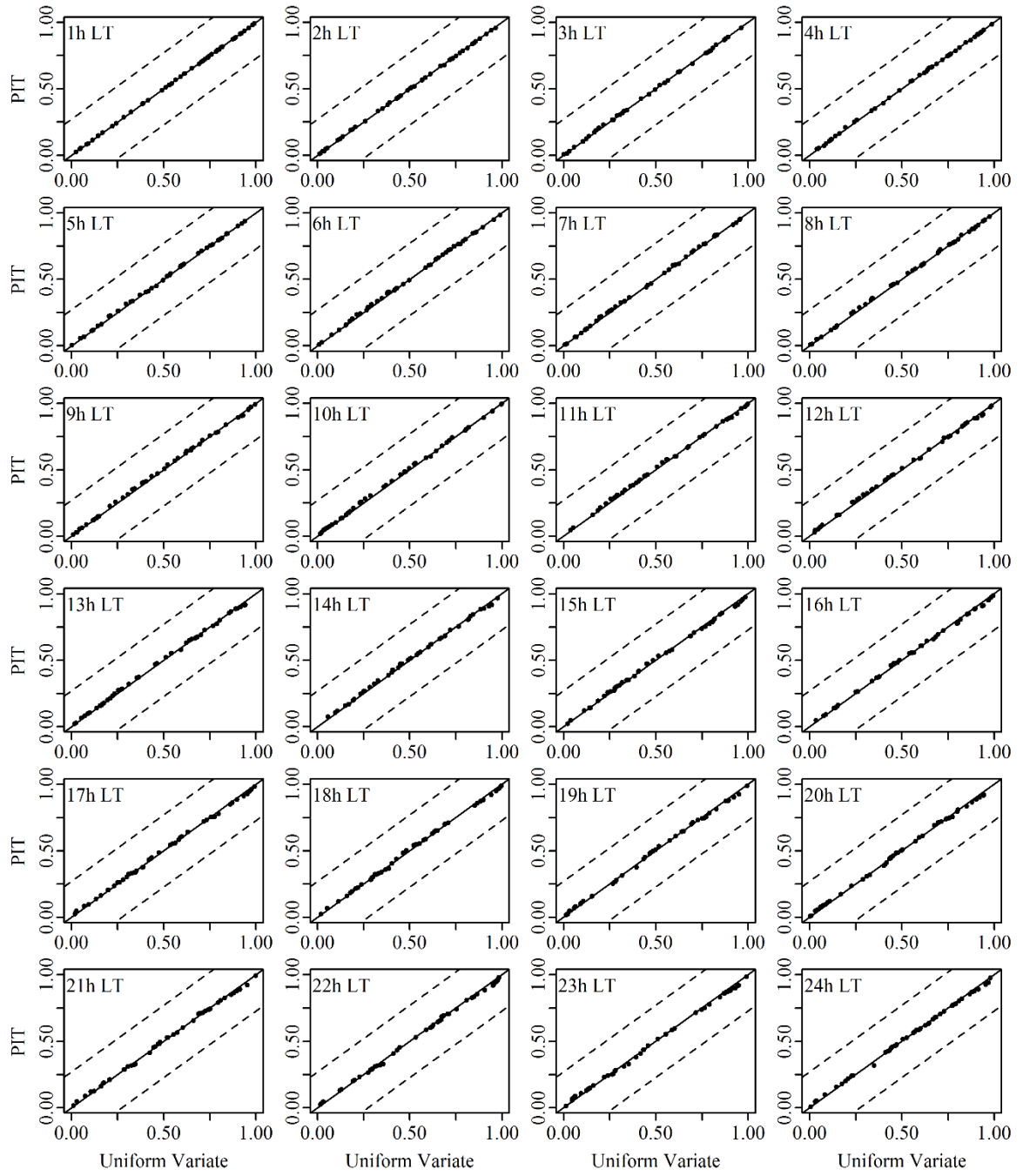


Figure A8. PIT uniform probability plot (autumn set; GML approach).

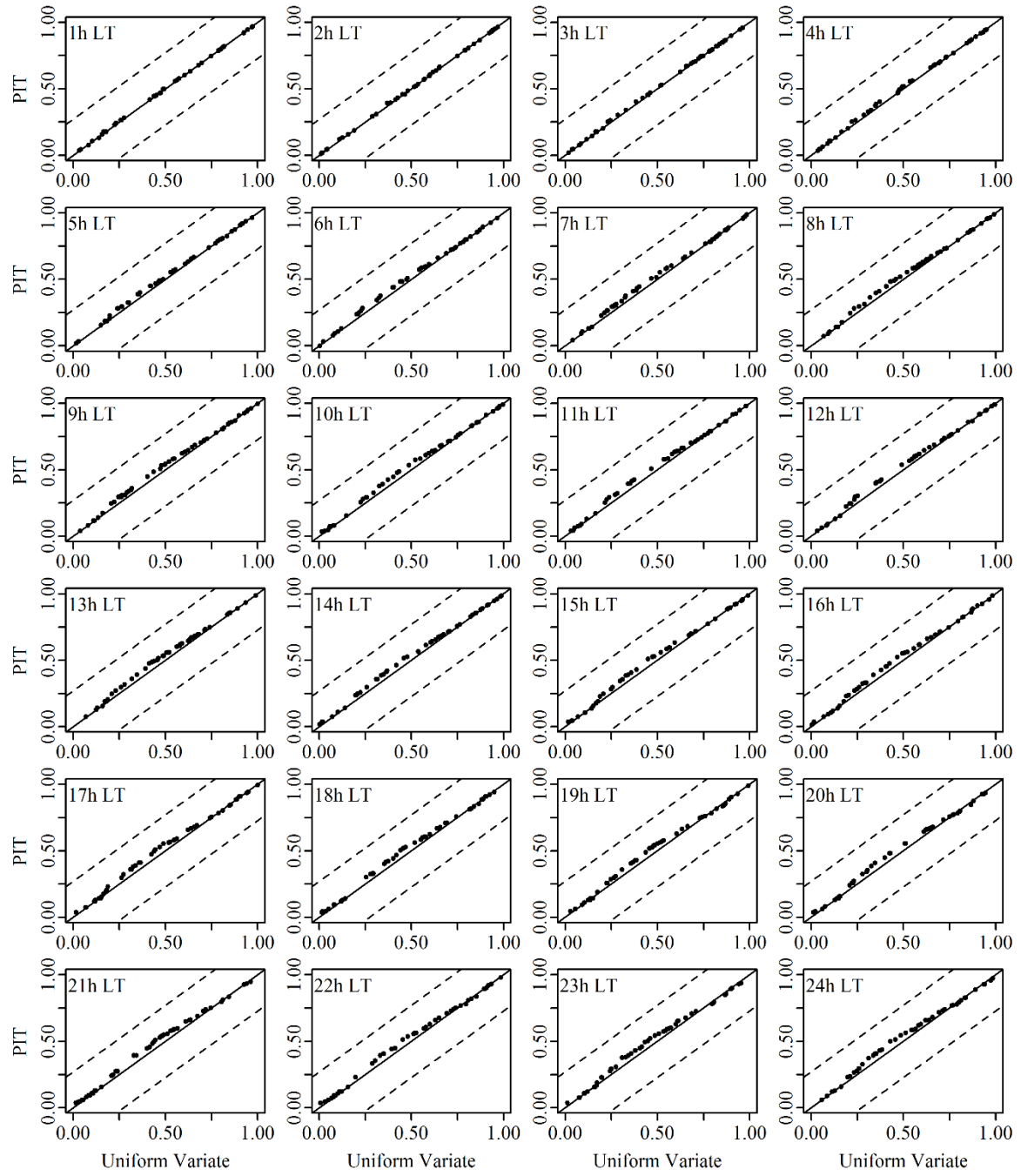


Figure A9. PIT uniform probability plot (winter set; GML approach).

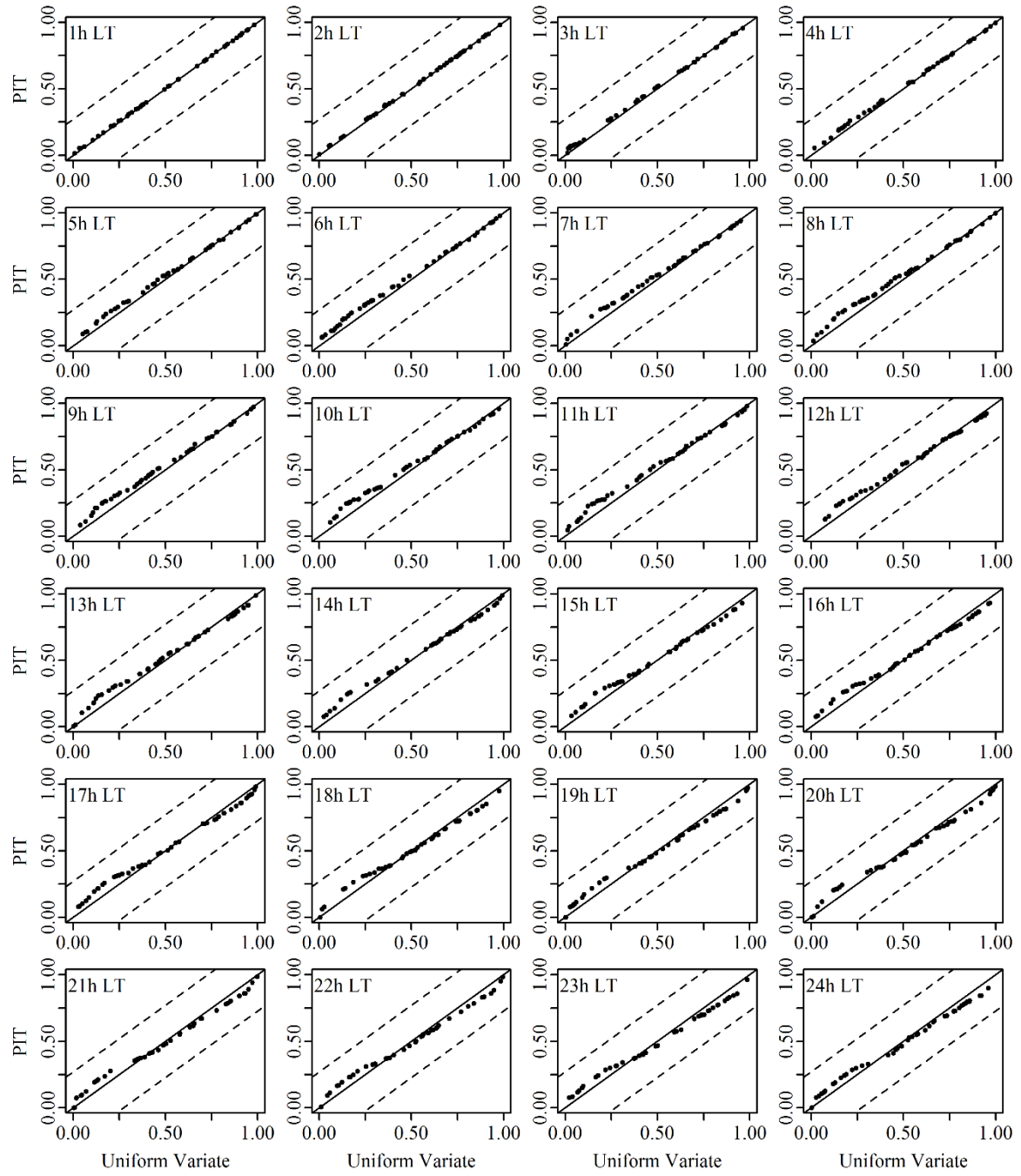


Figure A10. PIT uniform probability plot (spring set; GML approach).



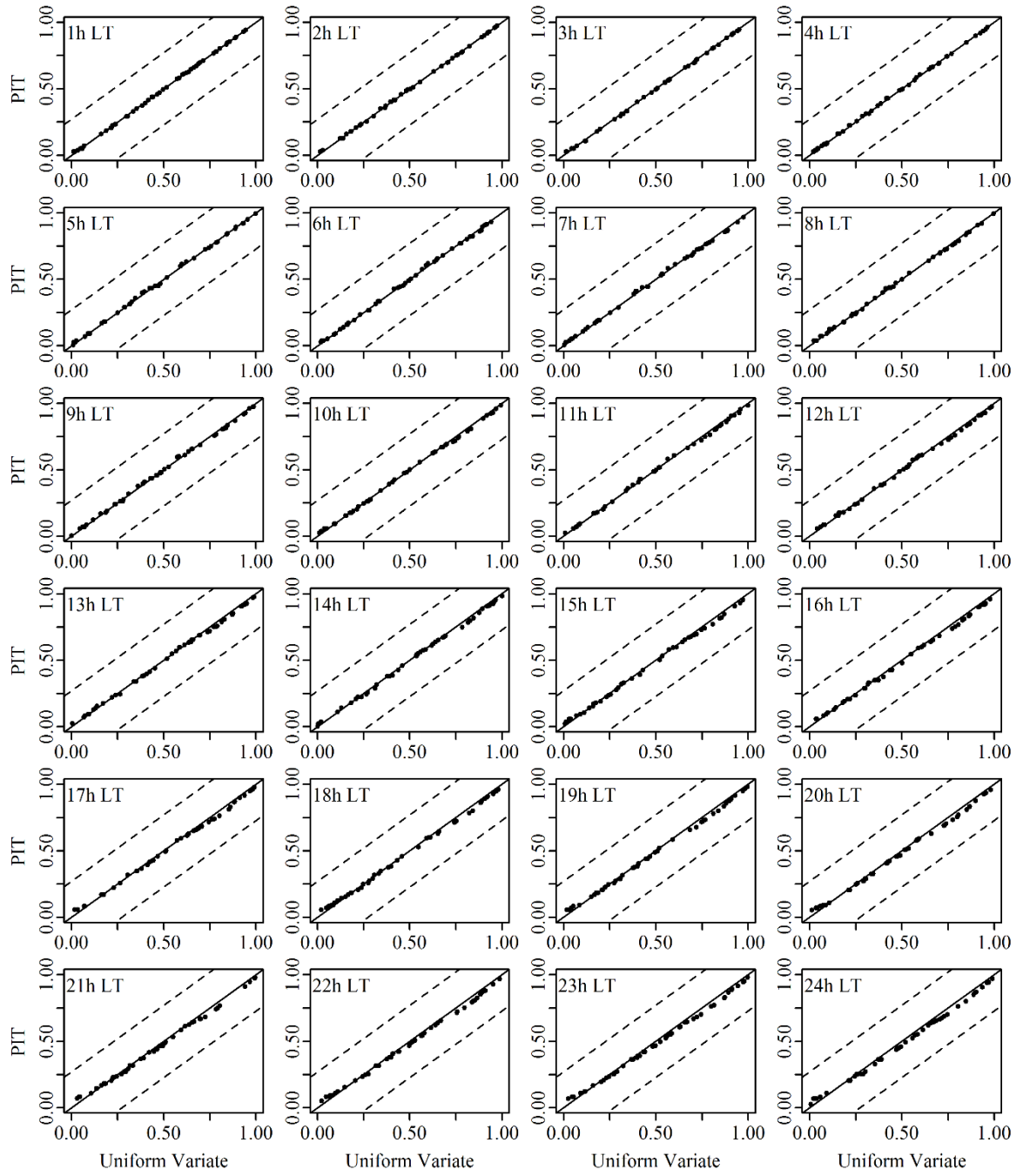


Figure A11. PIT uniform probability plot (summer set; GML approach).

# Improving real-time inflow forecasting into hydropower reservoirs through a complementary modelling framework

A. S. Gragne<sup>1</sup>, A. Sharma<sup>2</sup>, R. Mehrotra<sup>2</sup> and K. Alfredsen<sup>1</sup>

[1]{Department of Hydraulic and Environmental Engineering, Norwegian University of Science and Technology, Trondheim, Norway}

[2]{School of Civil and Environmental Engineering, The University of New South Wales, Sydney, Australia}

## Abstract

Accuracy of reservoir inflow forecasts is instrumental for maximizing the value of water resources and benefits gained through hydropower generation. Improving hourly reservoir inflow forecasts over a 24 hour lead-time is considered within the day-ahead (Elspot) market of the Nordic exchange market. A complementary modelling framework presents an approach for improving real-time forecasting without needing to modify the pre-existing forecasting model, but instead formulating an independent additive or complementary model that captures the structure the existing operational model may be missing. We present here application of this principle for issuing improved hourly inflow forecasts into hydropower reservoirs over extended lead-times, and the parameter estimation procedure reformulated to deal with bias, persistence and ~~hetroseedasticity~~heteroscedasticity. The procedure presented comprises an error model added on top of an un-alterable constant parameter conceptual model, the models being demonstrated with reference to the 207 km<sup>2</sup> Krinsvatn catchment in central Norway. The structure of the error model is established based on attributes of the residual time series from the conceptual model. Besides improving forecast skills of operational models, the approach estimates the uncertainty in the complementary model structure and produces probabilistic inflow forecasts that entrain suitable information for reducing uncertainty in the decision-making processes in hydropower systems operation. Deterministic and probabilistic evaluations revealed an overall significant improvement in forecast accuracy for lead-times up to 17 hours. Season based evaluations indicated that the improvement in inflow forecasts varies across

seasons and inflow forecasts in autumn and spring are less successful with the 95% prediction interval bracketing less than 95% of the observations for lead-times beyond 17 hours.

## **1 Introduction**

Hydrologic models can deliver information useful for management of natural resources and natural hazards (Beven, 2009). They are important components of hydropower planning and operation schemes where it is essential to estimate future reservoir inflows and quantify the water available for power production on a daily basis. The identification and representation of the significant responses of hydrologic systems have been diverse among hydrologists. Different hydrologists have incorporated their perceptions of the functioning of hydrologic systems into their models and come up with several rival models; some of them process based and others data-based (for thorough reviews of the historic development of hydrologic modelling refer to Todini, 2007 and Beven, 2012). These models can be grouped in to two main classes, conceptual and data-driven models.

Lumped conceptual hydrologic models are the most commonly used models in operational forecasting. Models of this class use sets of mathematical expressions to provide a simplified generalization of the complex natural processes of the hydrologic systems in the headwater areas of reservoirs. Application of such models conventionally requires estimating the model parameters by conditioning to observed hydrologic data. Unlike conceptual models, data-driven models establish mathematical relationship between input and output data without any explicit attempt to represent the physical processes of the hydrologic system. Reconciling the two modelling approaches and combining the advantages of both approaches (Todini, 2007), has produced some example applications in forecasting systems where the two modelling approaches are harmoniously used for improving reliability of hydrologic model outputs (e.g. Abebe and Price, 2003 and Solomatine and Shrestha, 2009).

Usefulness of a model for operational prediction is determined by the level of accuracy to which the model reproduces observed hydrologic behaviour of the study area. In operational applications, evaluation of how well the models capture rainfall-runoff processes, especially the snow accumulation and melting process in cold regions, is important because the extent to which the models accurately reproduce the reservoir inflows can significantly influence the efficiency of the hydropower reservoir operation and subsequently the power price. Application of hydrologic models for reproducing historic records can suffer from inadequacy in model

1 structure, incorrect model parameters, or erroneous data. Consequently, despite failing to  
2 reproduce the observed hydrographs exactly, they enable simulation of hydrologic  
3 characteristics of a study catchment to a fair degree of accuracy. It gets more challenging when  
4 using the models in the operational setup for forecasting the unknown future just based on the  
5 known past, which the model might not capture accurately. In the context of the Norwegian  
6 hydropower systems, being unable to predict future reservoir inflows accurately has negative  
7 consequences to the power producers. Norway's energy producers have to pledge the amount  
8 of energy they produce for next 24 hours in the day-ahead market and if unable to provide the  
9 pledged amount of energy the chance of incurring losses is very high. Estimation of future  
10 reservoir inflows (be it long- or short-term) involves estimating the actual (initial) state of the  
11 basin, forecasting the basin inputs during the lead-time, and describing the water movement  
12 during the lead-time (Moll, 1983). Hence, the quality of a hydrologic forecast depends on the  
13 accuracy achieved and methodology selected in implementing each of these aspects.

14 In this study, we intend to use conceptual and data-driven models complementarily. A  
15 conceptual model with calibrated model parameters is used as the fundamental model that  
16 approximately captures dominant hydrologic processes and forecasts behaviour of the  
17 catchment deterministically. A data-driven model is then formulated on the residuals, the  
18 difference between observations and predictions from the conceptual model. By studying the  
19 whole set of residuals and exploring the information they contain, important information that  
20 describes the inadequacies of the conceptual model can be extracted. In general, this kind of  
21 information can be used for improving either the conceptual model itself or the prediction skill  
22 of a forecasting system. Emulating the practice in most Norwegian hydropower reservoir  
23 operators, we stick to the latter purpose with the aim of enhancing the performance of a  
24 hydropower reservoir inflow forecasting system. According to Kachroo (1992), data-driven  
25 models defined on the residuals from a conceptual model can expose whether the conceptual  
26 model is adequate to identify essential relationships exhibited in the input-output data series.  
27 Data-driven models can establish the mathematical relationship that describes the persistence  
28 revealed in the residual time series, which is caused by failure of the conceptual model to  
29 capture all the physical processes exactly. Thus, in the operational sense, the data driven models  
30 can play a complementary role by adjusting output of the conceptual model whenever the  
31 conceptual model needs corrective adaptation (e.g. Serban and Askew, 1991 and  
32 World Meteorological Organization, 1992).

Several example applications can be found in the scientific literature on using conceptual and data driven models complementarily. For instance, Toth et al. (1999) compared performance improvements six ARIMA based error models brought to streamflow forecasts from a conceptual model to identify the best error model and data requirements. Shamseldin and O'Connor (2001) coupled a multi-layer neural network model on top of a conceptual rainfall-runoff model to improve accuracy of stream flow forecasts without interfering with operation of the conceptual model. Similarly, Madsen and Skotner (2005) developed a procedure for improving operational flood forecasts by combining error models (linear and non-linear) and a general filtering technique. Xiong and O'Connor (2002) investigated performance of four error-forecast models namely, the single autoregressive, the autoregressive threshold, the fuzzy autoregressive threshold and the artificial neural network updating models, for improving real-time flow forecasts and compared their results. Likewise, Goswami et al. (2005) examined the forecasting skill of eight error-modelling based updating methods. A recent review on the application of error models and other data assimilation approaches for updating flow forecasts from conceptual models can be found in Liu et al. (2012).

As reviewed above, the principle of complementing conceptual models with data-driven models has enjoyed applications in real-time hydrologic forecasting since the 1990s. The methodological contribution of the present work is reformulation of the parameter estimation procedure for the data-based model. We recognize that the bias, persistence and heteroscedasticity seen in the residuals from the conceptual model reflect structural inadequacy of the conceptual model to capture the catchment processes and, hence, are important in defining the manner the residual series is dealt with. Accordingly, we describe the reservoir inflows in a transformed space and present an iterative algorithm for estimating parameters of the data-driven model and the transformation parameters jointly.

Two main features distinguish application aspects of the present paper from previous published works built on the same concept of complementing conceptual models with data driven models. Firstly, it attempts to provide hourly reservoir inflows of improved accuracy 24 hours ahead. The earlier papers mainly succeeded in improving forecasts for forecast lead-times up to six time steps or incorporated a scheme to update the forecast system at an interval of six time-steps. Secondly, an attempt is made in what follows, to produce a probabilistic forecast by estimating the uncertainty of the error model, rather than only the deterministic estimate. This, thereby, enables forecast of an ensemble of reservoir inflows, thereby allowing a risk-based

paradigm for hydropower generation being put to use. Reasons as to why hydrologic forecasts should be probabilistic, and the potential benefits therein are presented and explained in Krzysztofowicz (2001). Krzysztofowicz (1999) describes a methodology for probabilistic forecasting via a deterministic hydrologic model. Li et al. (2013) provide review of scientific papers that provide various regression and probabilistic approaches for assessing performance of hydrologic models during calibration and uncertainty assessment. Smith et al. (2012) demonstrate a good example of producing probabilistic forecasts based on deterministic forecast outputs. In this paper, the improvement levels achieved are evaluated deterministically using the same or similar metrics as past studies, and probabilistically using: (i) the containing ratio (Xiong et al., 2009), which is also referred to as reliability score (e.g. Renard et al., 2010-2010); and (ii) the probability integral transform (PIT) plot, which Thyer et al. (2009) refer to as the predictive QQ plot. We here emphasise that taking into account uncertainties emanating from various recognized sources and describing the degree of reliability of the inflow forecasts has important benefits. According to Montanari and Brath (2004), the Bayesian forecasting system (BFS) and the generalized likelihood uncertainty estimation (GLUE) are the popular methods for inferring the uncertainty in hydrologic modelling. Yet, the scope of producing probabilistic inflow forecasts in this study is limited to attaching a certain probability to the deterministic forecasts so common in the Norwegian hydropower industry based on analysis of the statistical properties of the error series from the conceptual model, and assessing its degree of reliability.

In the next section, the complementary model setup is formulated and the performance evaluation criteria are provided. An example application is presented in the subsequent section. This includes description of the study area and data used, findings from the evaluation of the complimentary setup and its components during calibration and validation, and results of forecasting skill assessment using deterministic and reliability metrics. Finally, concluding remarks are provided.

## 2 Methodology

### 2.1 The conceptual model setup

The widely applied conceptual hydrologic model—HBV—(Bergström, 1995) is used in this study. The version used allows dividing the study catchment up to 10 elevation zones. A

deterministic HBV model with already calibrated model parameter values was assumed to take the role of the operational hydrologic models Norwegian hydropower companies commonly use for forecasting reservoir inflows. In the operational setup, the air temperature and precipitation input over the forecast lead-time are obtained from the Norwegian Meteorological Institute ([www.met.no](http://www.met.no)). As this study aims to improve hydrologic forecasts into the hydropower reservoirs by complementing the conceptual model by an error model, we assume that the predictions from the HBV model are made using as good quality input data as possible. Hence, the observed air temperature and precipitation data are used as input forecasts in hindcast.

## 2.2 The complementary error model

The error model aims at exploiting the bias, persistence and heteroscedasticity in the residuals and estimating the errors likely to occur in the forecast lead-time. Forecasting the error in the lead-time is regarded as a two-step process: off-line identification and estimation of the error model, and error predictions based on most recent information.

### 2.2.1 Identification of the model structure

An error model that captures the structures the processes model is missing should lead to a zero-mean-homoscedastic residual series from the modelling framework. In order to identify the right structure and establish a parsimonious model that adequately describes the data, we diagnose the residuals and address the bias, persistence and heteroscedasticity the series might exhibit as follows.

First and foremost, we transform the observed ( $Q$ ) and the predicted ( $\hat{q}$ , from the conceptual model) inflows into  $z$  and  $\hat{z}$ , respectively. This way we deal with the heteroscedasticity seen in the residuals by making repeated use of Eq. 1 with the appropriate inflow term.

$$\hat{z}_t = \begin{cases} \left( (\hat{q}_t + \beta)^\lambda - \beta \right) \lambda^{-1} & \lambda > 0 \\ \log(\hat{q}_t + \beta) & \lambda = 0 \end{cases} \quad (1)$$

where  $\beta$  and  $\lambda$  are the transformation parameters.

The discrepancy ( $\varepsilon$ ) between the observed and predicted inflow at time step ( $t$ ) can be expressed as  $\varepsilon_t = z_t - \hat{z}_t$ . Analysis of whether the residuals are random or show some bias

follows. Lest the mean of the residuals would be different from zero, the mean error ( $\mu_e$ ) is subtracted from the error series ( $\varepsilon$ ) to produce a zero-mean residual series ( $e_t = \varepsilon_t - \mu_e$ ). This is followed by assessment of the auto correlation function (acf) and partial autocorrelation function (pacf), which are keys for identifying the order of Markovian dependence the residuals exhibit. We consider an autoregressive (AR) model structure (Eq. 2) to represent the persistence structure in the residual series. Comparative assessment of error models of different complexity would be an interesting work but is beyond the scope of this study. Xiong and O'Connor (2002) affirm that AR model's longstanding popularity is deservedly right and further emphasize effectiveness of a very parsimonious model such as AR model for error forecasting.

$$\hat{e}_t = \sum_i^p a_i e_{t-i} \quad (2)$$

where  $p$  designates the length of the lag-time, and  $a_1, a_2, \dots, a_p$  are coefficients of the AR model.

In order to provide improved hourly reservoir inflow forecasts over a 24 hours lead-time, the error-forecasting model takes the form of Eq. (3). In order to overcome lack of observed residuals encountered for forecast lead-time ( $f$ ) longer than one-step ahead, it is necessary to utilize estimated errors as inputs (see Eq. 3). The number of estimated errors values to be used as inputs depends on the identified order of the AR model and can vary across the forecast lead-times.

$$\hat{e}_{t+f} = \begin{cases} \sum_{i=1}^p a_i e_{t+f-i} & \text{for } f = 1 \\ \sum_{i=1}^{f-1} a_i \hat{e}_{t+f-i} + \sum_{i=f}^p a_i e_{t+f-i} & \text{for } f \geq 2 \text{ and } p \geq f \\ \sum_{i=1}^p a_i \hat{e}_{t+f-i} & \text{for } f \geq 2 \text{ and } p < f \end{cases} \quad (3)$$

In its complete form, the error-corrected reservoir inflow forecast ( $z'$ ) from the complementary modelling framework can be given as

$$z'_{t+f} = \hat{z}_{t+f} + (\mu_e + \hat{e}_{t+f}) \quad (4)$$



## 2.2.2 Parameter Estimation

Parameters of the AR model can be set to the corresponding Yule-Walker estimates of  $a_1, a_2, \dots, a_p$  given the autocorrelation function of the error series fulfils a form of linear difference equation. However, in practice, Eq. (2) can be treated as a linear regression and parameters can be estimated by Least Squares method as demonstrated by Xiong and O'Connor (2002). An iterative algorithm suggested in Beven et al. (2008) is adopted for estimating the model parameters while optimizing transformation of the inflow data. Adoption of a methodology that amalgamates parameter estimation and Box-Cox (Box and Cox, 1964) inspired transformation of inflow is useful for taking into account the heteroscedastic residuals and obtaining a normally distributed residual series from the error model. The parameter and inflow transformation steps with a little modification from Beven et al. (2008) over the calibration period  $(1, \dots, T)$  are as follows:

1. Select values of  $\beta, \lambda > 0$  and transform the reservoir inflows  $(\hat{q}_{1:T}, Q_{1:T})$  to get  $(\hat{z}_{1:T}, z_{1:T})$  using Eq. 1.
2. Calculate the residuals series from the transformed inflow data  $(\varepsilon_{1:T} = z_{1:T} - \hat{z}_{1:T})$ .
3. Perform an optimization for the error model parameters  $(a_1, a_2, \dots, a_p)$  to minimize  $\sum (\varepsilon_{1:T} - \hat{\varepsilon}_{1:T})^2$ , where  $\hat{\varepsilon}$  represents the forecast from the error model which at a given observation time step  $(t)$  equals  $(\mu_e + \hat{e}_t)$ . Thus, the observed  $(\varepsilon)$  and forecasted  $(\hat{\varepsilon})$  errors at a given observation time step  $(t)$  can be related as  $\varepsilon_t = \hat{\varepsilon}_t + \eta_t$ , where  $\eta_t$  is a random noise that describes the total uncertainty originating from various sources.
4. Adjust  $(\beta, \lambda)$  and repeat the optimization until the residuals of the error model appear homoscedastic. The  $\eta_t$  term (step 3) is assumed to be unimodal, symmetric and unbounded random variable with a zero expected-mean and second moment given as  $\sigma^2$ .

## 2.3 Performance evaluation

In addition to visual evaluation of the hydrographs, performance of the present procedure is robustly analysed using deterministic and reliability metrics. The root mean square error

(*RMSE*), relative error (*RE*) and the Nash-Sutcliffe efficiency (*NSE*) (Nash and Sutcliffe, 1970) are employed to evaluate efficiency of the models during calibration and validation deterministically. Evaluations are made with respect to varying forecast lead-times and season wise as well. Among the three statistical performance criteria, the *RE* (Eq. 5) measures the relative error between the total observed and predicted inflow volume. For a good simulation the value of *RE* is expected to be close to zero. Quantifying the relative error (*RE*) of the simulations/forecasts is important because it indicates how the inaccuracies affect a hydropower company's ability to deliver the amount of energy it has pledged to provide to the energy market. Therefore, special attention is given to the less aggregate version of *RE*, which we hereon refer to as percentage volume error (*PVE*) and describe as follows.

$$RE = \frac{\sum (z_t - \hat{z}_t)}{\sum z_t} \times 100\% \quad (5)$$

The *PVE* designates the relative error at each time step, which in reference to Eq. 5 can be obtained by omitting aggregation of the errors by summation. It indicates the magnitude of the errors as percentage of the observed inflows at each inflow time step. From hydropower systems operations point of view, the *PVE* enables evaluation of the forecast errors at each time step and assess implication on the power production capacity directly. The *PVE* analysis devised here divides the computed *PVE*s into six *PVE* classes (i.e.  $\leq 10\%$ , 10-20%, 20-30%, 30-40%, 40-50% and  $>50\%$ ), and treats overestimates and underestimates separately. The number of times each of the six absolute *PVE* classes appeared in the set or subset of interest (i.e. hydrologic year or seasons) is constructed by keeping score of the *PVE* class into which each and every residual fell in. Then the fraction of time each *PVE* class occurred is divided to the total number of points in the given set/subset and is reported as a percentage. This is designated as a “*PVE* count”. Model performance assessment using *PVE* (during simulation and forecasting) mainly focuses on assessing the change in number the number of incidences in each *PVE* set, which in other words means the change in *PVE* counts. The *PVE* count/change in *PVE* count, along with the above-mentioned deterministic statistical criteria, is used for evaluating simulation and forecasting skill of the complementarily setup system (conceptual model + error model). As a metric for measuring relative improvement in forecasting skills, high *PVE* counts for the low *PVE* classes (e.g.  $\leq 10\%$ ) is considered desirable quality. The justification is that, the penalty a power producer incurs when failing to deliver the pledged

amount of power would be lesser if its forecasting system makes errors of lower PVE classes more frequently.

Another useful metric used for assessing forecasting skill of the complementary setup is through uncertainty analysis. An interval forecast (Chatfield, 2000) can be constructed by specifying an upper and lower limit between which the future reservoir inflow is expected to lie with a certain probability  $(1-\alpha)$ . The prediction interval for the inflow forecast are estimated using the Linear Regression Variance Estimator (LRVE) Shrestha and Solomatine (2006) describe. Xiong et al. (2009) outline several indices that can serve for describing the properties of prediction bounds of particular probability and for comparative study of prediction intervals resulting from different uncertainty assessment schemes. The indices characterise the prediction bound either by: the percentage of observations it contains, its band-width, or its symmetry relative to the observation. According to Xiong et al. (2009), of all ~~indeces~~indices the containing ratio (*CR*), which describes the percentage of observed inflows falling in the desired interval percentage, is the widely used metrics for assessing reliability of probabilistic forecasts. We adopt the *CR* metric for describing the reliability of the forecasts with the desired interval percentage of 95% ( $\alpha = 0.05$ ). Beside the *CR*, we verify the probabilistic forecasts graphically using the less formal PIT uniform probability plot. The working procedure along detailed application examples can be found in Laio and Tamea (2007) and Thyer et al. (2009). Among others, Pokhrel et al. (2013) and Wang et al. (2009) demonstrate viability of the ‘PIT uniform probability plot’ approach for checking uniformity (and investigating the causes, in cases of deviations from uniformity) without binning the data subjectively.

### 3 Example application

#### 3.1 Study area and data

The Krinsvatn catchment is located in Nord Trøndelag County in mid-north Norway. It comprises an area of 207 km<sup>2</sup> and about 57% of the catchment is mountain area above timberline. The elevation ranges from 87 to 628 m above mean sea level and is drained by the Stjørna/Nord River. The dominant land use is forest covering 20.2% of the study site while marsh, lakes and farmlands cover about 9%, 6.7% and 0.4% of the catchment area, respectively. Figure 1 provides location and main characteristics of the study site, and the daily potential evapotranspiration values used.

Observed hourly data of eleven water-years (2000/01 to 2010/11) was split into three sets used for warming-up (2000/01), calibrating (2001/02-2005/06) and validating (2006/07-2010/11) the conceptual and the error models alike. Observed precipitation and temperature data of two meteorological stations (i.e. Svar-Sliper and Mørre-Breivoll) in neighbouring catchments are used. Discharge data for the catchment is derived from water level records at the Krinsvatn gauge station. Beven (2001) outlines the advantages to direct use of water level information in hydrologic forecasting. Rating curve uncertainties and their influence on the accuracy of flood predictions have been documented very well (e.g. Sikorska et al. 2013; Aronica et al., 2006; Pappenberger et al. 2006; Petersen-Overleir et al. 2009). Krinsvatn is considered a stable discharge measurement site with few external influences, and the rating curve was updated in 2004. This study, however, considers the uncertainty of the rating-curve to be one of the factors contributing to the total error expressed in Eq. 2 and does not address it separately.

### 3.2 HBV model for Krinsvatn catchment

The catchment is divided into 10 elevation zones in the HBV model setup. Input data used are hourly areal precipitation, air temperature, and potential evapotranspiration. The model is run on an hourly time step for water years 2000/01 to 2005/06 with the last five water years being used for model calibration. Calibration is carried out using the shuffled complex evolution algorithm (Duan et al., 1993), with the *NSE* between the observed and predicted flows as an objective function. Description of the model parameters along the corresponding optimized values is provided in Table1.

#### 3.2.1 Overview of the conceptual model's performance

The simulation and observed reservoir inflow hydrographs shown in Fig. 2 indicate a certain level of agreement for most of the calibration and validation periods, which the statistical evaluations (Table 2) agree with. The overall hourly reservoir inflow predictions during calibration and validation show efficiency of  $NSE > 0.5$  and  $RE < \pm 25\%$ ; even though simulations match observations better during calibration than validation. High *NSE* values ( $> 0.8$ ) during both calibration and validation reveal that the inflow simulations fit the observed hydrographs best in the winter seasons. Nevertheless, it is evident that model predictions in the validation period are prone to underestimation bias ( $RE > 0$ ). Season wise assessment of the validation period reveals the conceptual model's tendency to underestimate reservoir inflows in spring and summer considerably. In light of what the *NSE* and *PBISRE* metrics suggest, the

1 lower RMSE values (i.e. for instance summer season) do not reflect superior model  
2 performances.

3 PVE counts of the six PVE classes (i.e.  $\leq 10\%$ , 10-20%, 20-30%, 30-40%, 40-50% and  $>50\%$ )  
4 are computed on the residuals between observed and simulated reservoir inflows. The stacked-  
5 columns of Fig. 3a&b show how frequently each of the six absolute PVE classes occurred over  
6 the calibration and validation period. The results reveal a large degree of discrepancy between  
7 observations and predictions during calibration and validation. Simulated inflows deviated from  
8 the corresponding observed values by a magnitude of more than  $\pm 10\%$  in about 83.3%  
9 (calibration) and 88.6% (validation) of the respective simulation time steps. Huge difference  
10 between observations and simulations is noted in the summer season with absolute PVE of the  
11 class  $>50\%$  occurring in more than half of the simulation time steps throughout the calibration  
12 and validation periods. Winter simulations listed the highest level of occurrence of PVE of the  
13 class  $\leq \pm 10\%$  during both calibration and validation. Comparable to the results in Table 2,  
14 volume errors in winter simulations do not seem to be a serious problem, probably because the  
15 season is predominantly a snow accumulation rather than runoff generation period. Errors of  
16 the high absolute PVE classes scored high PVE counts in the spring and autumn seasons.

17 Details of the extent to which the reservoir inflows are under- and over-estimated can be seen  
18 in Fig. 3c&d. The fraction of time the simulated inflows exhibited under- and over-estimation  
19 during calibration is 51.9% and 46.8%, respectively. In the validation period, the reservoir  
20 inflows are underestimated about 65.6% of the time compared to overestimation in 33.4% of  
21 the times. This is also revealed in the findings from statistical metrics in Table 2, which disclose  
22 the bias in the model. Yet, the results in Fig. 3 further reveal that the model predictions deviate  
23 from the observations at high discharges. For example, during the validation period 59.2% of  
24 the times observations exceeded the predictions by magnitudes more than 10%. Such  
25 information is useful because direct evaluation of observed and predicted values explains the  
26 implications of model performance on the planning and operation of a hydropower system  
27 better than an aggregated variance based statistic. From an operational management point of  
28 view, considerable underestimation of reservoir inflows can have both short- and long-term  
29 effects on the operation of a hydropower system. In the short-term, the company could be forced  
30 to release unvalued water especially when the reservoir water level is close to its maximum  
31 capacity. Hence, the high percentage of underestimations that occur in the autumn and spring  
32 seasons (during calibration and validation) should not be tolerated because the inflows in the

autumn and spring seasons are very important. On the one hand, substantial overestimation of reservoir inflows can at least expose any Norwegian hydropower company to undesirable expense due to obligations to match the power supply it has failed to deliver by dealing with other producers in the intra-day physical market (Elbas). Although overestimation does not seem to be a pertinent issue, Fig. 3d unmasks that the inflows are overestimated by a magnitude >50% at least 10% of the time in all seasons.

### 3.2.2 Residual analysis

Following the example of Xu (2001), a Kolmogorov-Smirnov test is applied to residuals of the conceptual model. The test revealed that the residuals are not normally distributed. The maximum deviation between the theoretical and the sample lines is 0.130, which is larger than Kolmogorov-Smirnov test statistic of 0.008 at significance level  $\alpha = 0.05$ .

Presence of homoscedasticity in the residuals series is diagnosed visually by plotting the residuals versus the predicted reservoir inflows (Fig. 4a). With respect to the horizontal axis, the scattergram does not remain symmetric for the entire range of predicted inflows. The residuals show high variability and possible systematic bias when inflows are less than 3.5mm while the opposite is true when the inflows exceed 3.5mm. Inflows of magnitudes between 3.5 and 5.5mm seem to be underestimated while overestimation is visible when the inflow rates are greater than 5.5mm. However, as can be seen from Fig. 2, inflows of magnitude up to 3mm represent reservoir inflows during the rise of the hydrographs including all peak inflows for all hydrologic years but 2005/2006 and 2010/2011. Hence, except for the possible systematic bias during low flows, the inference from the scatterplot is inconclusive to support or dismiss the issue of predominant underestimation revealed in the model performance evaluation. Moreover, hourly inflows of magnitudes higher than 3mm are rare and occurred about 0.1% of the times over the calibration and validation period.

Plots of autocorrelation and partial autocorrelation functions of the residual time series (Fig. 4b&c) indicate a strong time persistence structure in the error series. Rapid decaying of the partial autocorrelation function confirms the dominance of an autoregressive process, which the gradually decaying pattern of the autocorrelation function also suggests. Thus, in order to obtain a Gaussian series it is important to address issues of heteroscedasticity and serial correlation in the residual series. As the current study aims at utilising the persistent structure in the residuals for supplementing the forecasting system, the corrective action to be taken only

aims at removing the heteroscedasticity. A successful way to do it is through transformation of the flow data (e.g. Engeland et al., 2005). As outlined in the methodology section, the reservoir inflows (both observed and predicted) are transformed while estimating parameters of the error model.

### 3.3 Structure and performance of the error model

In accordance with the findings from the ACF and PACF plots discussed in section 3.3.2, AR models of up to order  $p = 3$  were investigated while estimating parameters of the error model. As outlined in section 2.2.2, coefficient of the  $AR(p)$  model and the transformation parameters were estimated by minimizing the sum of the squares of the offsets between the inflows (observed and predicted) in the transformed space, and assessment of whether the subsequent residuals from the complementary modelling framework appear homoscedastic and exhibited correlation. The latter was assessed using the Kolmogorov-Smirnov ( $KS$ ) statistic as a relative quantitative measure followed by visual inspection of the residual plots, which led to the selection of an  $AR(1)$  model with transformation parameters  $\beta = 41.4$  and  $\lambda = 0.9$ , bias correction  $\mu_e = 0.021$  and coefficient  $a_1 = 0.97$ .

Calibration efficiencies calculated for the error model using the  $RMSE$ ,  $RE$  and  $NSE$  metrics are 0.096, -100% and 0.517, respectively. Corresponding values for the validation period are computed as 0.095, 20.3% and 0.630, respectively.  $NSE$  values for the calibration and validation periods imply ability of the error model to capture at least half of the discrepancies observed between observations and predictions from the conceptual model. The All the three metrics reveal a higher efficiency in the validation set than the calibration set. With reference to Table 2, this suggests too much fitting of the HBV model to the data that led to extraction of more information from the calibration set. Assessment of the residuals from the complementary framework reveals that the transformation reduced the maximum deviation between the theoretical and the sample lines slightly from 0.13 to 0.10; yet the residuals are not normally distributed (i.e. Kolmogorov-Smirnov statistic of 0.008 at significance level of  $\alpha = 0.05$ ). This implies that the assumption the residuals from the complementary forecasting system would be Gaussian is far from being true. As the aim of this study is to utilize the error and complementary models additively, we discuss in the next section the extent to which the complementary setup boosted prediction ability in the forecasting mode ~~is discussed in the next~~

~~section~~ and come back to the issue of violation of the Gaussian assumption in section 3.5, where we analyse the reliability of the forecasts probabilistically.

### 3.4 Forecasting skill of the complementary setup (deterministic assessment)

Imitating operational application of forecasting models in the Norwegian hydropower system, reservoir inflows for the day-ahead market (Elspot) are estimated using the presented forecasting system. The system has to run once a day at an hourly time step, sometime before 12 pm after retrieving the latest observations, and the inflow forecasts are issued for the next 24 hourly time steps beginning from 12 o'clock noon. Overall performance of the complementary model in forecasting the reservoir inflows during the calibration and validation periods is first discussed and is followed by evaluation of its forecasting skill with respect to forecast lead-times. Evaluation of the forecast skill presented in this paper is based on assessment of forecasts made for the period between 2006/07 and 2010/11 as the datasets from 2000/01 to 2005/06 are used for calibrating the system.

#### 3.4.1 Overall performance

Assessment of the overall forecasting skill of the complementary setup shows significant improvement in forecast accuracy. The *RMSE* and *NSE* statistical criteria computed between forecasted and observed inflows are 0.095 and 0.896, respectively. *RMSE* values for the autumn, winter, spring and summer forecasts are 0.094, 0.090, 0.132 and 0.044, respectively, and the corresponding *NSE* values are 0.904, 0.905, 0.859 and 0.873.

Proving capability of the complementary setup to reduce the bias revealed in the simulation forecasts from the conceptual model, which was pointed out in the previous section, the 24 hours lead-time forecasts exhibited low-level underestimation bias with RE equal to 3.8%. Degree of bias in the inflow forecasts differed seasonally. RE computed for each season in a decreasing order is, summer (-10.2%), spring (4.6%), autumn (2.9%) and winter (0.7%). The relatively higher bias in the spring and autumn forecasts can be related to runoff generation in the Krinsvatn catchment due to snow melting or occurrence of precipitation in the form of rainfall, which can affect the persistence structure in the residual series obtained from the conceptual model.

Stacked-column plots in Fig. 5 display the occurrence level of each of the six PVE classes in the residual series between forecasts and observations. Visual comparison of stacked-column



plots of Fig. 5 and Fig. 3 shows reduction in PVE count of the high PVE classes and increase in PVE counts of low PVE classes; e.g., PVE count for the PVE class  $>\pm 50\%$  decreased by about 15% while PVE count for the PVE class  $\leq \pm 10\%$  grew by about 50%. In order to assess this assertion, a further assessment is carried out by dividing the six PVE classes into two groups: low PVE ( $PVE \leq \pm 10\%$ ) and high PVE ( $PVE > \pm 10\%$ ). Ratio between seasonal PVE counts of the low and high PVE classes is taken and comparison is made on two sets of residual series. These sets of residuals are, (1) residuals from the simulated forecasts (conceptual model), and (2) residuals from forecasts of the complementary setup. Results are presented in Table 3. Apart from confirming the success in reducing PVE counts of high PVE errors, the results indicate that equal level of success is not achieved in all four seasons. In relative terms, high PVE errors occur more often in the spring and summer forecasts. As pointed out earlier, this can be associated to the snowmelt and, to a certain degree, to rainfall incidents occurring in these seasons.

### 3.4.2 Forecast skill with respect to forecast-lead times

Relative reductions in *RMSE* between forecasts from the complementary setup and the simulated forecasts from the conceptual model are computed. Detailed results for each season of the hydrologic years between 2006/07 and 2010/11 are presented in Table 4. The results are also summarized in terms of the minimum, mean and maximum relative *RMSE* reduction as shown in Fig. 6. Excluding forecasts in autumn and winter seasons of 2006/07, relative *RMSE* reductions are observed in forecasts of short and long lead-times. Of course, in all four seasons, the achieved level of improvement in forecast accuracy is high for short lead-times and diminishes gradually with increased lead-time. Results show that accuracy of the reservoir inflows in the spring and summer seasons are improved over the entire range of the forecast lead-time. Likewise, reduction in *RMSE* is observed for all autumn and winter inflow forecasts except for years 2006/07 and 2007/08, respectively.

In order to get insight on the improvement level in a unit directly related to hydropower production, the change in PVE count of each PVE class is calculated. Change in PVE count of a given absolute PVE classes is the difference between the PVE counts for the complementary setup and that for the conceptual model. The results are summarized as shown in Fig. 7. The figure shows that the PVE count of high magnitude absolute PVE classes are reduced and the opposite is true for that of the smaller absolute PVE classes. For instance, regardless of the type of discrepancy (under- or over-estimation) noted, the change in PVE counts of the absolute *PVE*

of the class  $>50\%$  is negative. The negative sign implies less errors falling in this PVE class in the residual series from the complementary setup than those from the conceptual model. Similarly, the changes in PVE counts of the 20-30%, 30-40% and 40-50% absolute PVE classes indicate lowered fraction of occurrence of errors of these orders. In both cases of under- and over-estimation, absolute *PVE* of the class  $\leq 10\%$  occurred more frequently; for example, the fraction of time reservoir inflow forecasts of 1 hour lead-time deviated from the observations by a magnitude  $\leq 10\%$  increased by about 52.7 and 27.7% during under- and over-estimations. Overall, the plots show that the magnitude of discrepancy at each forecasting point is significantly reduced. The improvement level at each forecast lead-time is proportional to the vertical distance from the horizontal axis. It can be noted that, the vertical distance narrows down with increasing lead-time suggesting a declining improvement level with increased lead-time.

Calculation of the relative RMSE reduction and the change in PVE counts agree that the forecast accuracy is improved through the complementary setup. The assessments further revealed that the degree of improvement weakens with increased forecast lead-time. However, the relative RMSE reduction computations indicate that in some occasions the simulated inflow forecasts stand out to be better. The relative RMSE reduction values for lead-times longer than 20 hours (Table 4) show that complementing the conceptual model with an error model is counterproductive in autumn and winter seasons of years 2007/08 and 2006/07, respectively.

### **3.5 Reliability of the inflow forecast**

Computation of the containing ratio (*CR*) for the entire forecast reveals that 96% of the observations are inside the 95% prediction interval. The inflow hydrographs (Fig. 8) confirm that most of the observed inflows are contained in the specified uncertainty bounds.

The percentage of observation points falling within the 95% prediction interval varies from season to season and across hydrologic years (see Fig. 9a). All observed winter and summer inflows are bracketed in the 95% uncertainty bound at least 95% of the time. In general, the winter season is more of a snow accumulation period and a closer observation of the hydrographs (see Fig. 8) reveals that the summer hydrographs cover the recession and base flow portions of the annual hydrographs. Thus, better persistence structure and predictable discrepancies between simulated forecasts from the conceptual model and the observations. As

Goswami et al. (2005) argue, the persistence structure in residual series primarily arises from the dynamic storage effects of a catchment system.

The desired percentage of autumn observations is contained in the 95% prediction interval in the years 2006/07, 2008/09 and 2010/11. In the years 2007/08 and 2009/10, however, only 93 and 94% of the observed autumn inflows are bracketed in the estimated 95% prediction intervals, respectively. Reliability score (*CR*) calculations for the spring season indicate that percentage of observation points falling in the desired prediction interval percentage are below 95% except in the hydrologic years 2007/08 and 2008/09. Unlike winter and summer inflows, autumn and spring flows mostly cover portions of the hydrograph corresponding to the rising limb or high flow regime (see Fig. 8). While physical factors contributing to the increase in quick flow into the reservoir are precipitation incidents (in the form of rainfall) and melting of snow in the headwaters, comprehension of this concept and its encapsulation into the HBV model leaves control of the catchment response to two threshold values (TX and TS, see Table 1 for description). Employing such simple threshold values to govern initiation of the runoff generation process based on air temperature measurement at a given time-step obviously involves more sources of uncertainty (i.e. measurement, model structure and model parameters). For instance, we assume the input air temperature at a given time step is erroneously recorded to be higher than TX and/or TS due to measurement error. Subsequently, the model will partition the precipitation as rainfall and initiate melting of snow, which the observation does not reveal. This kind of misclassification of precipitation and/or misrepresentation of snow accumulation and melting processes can simply occur due to the error in the input temperature record. Because of this, the persistence in the errors between simulated forecasts from the conceptual model and the observations can get weaker. According to Goswami et al. (2005), some degree of persistence in the model input (i.e. rainfall) is another primary source of the persistence characteristic of observed flow series. Even though the least *CR* calculated for the autumn and spring seasons are by no means too bad (i.e. 93% and 90%, respectively), the requirement for reliability is for the uncertainty bound to contain as much fraction of observations as desired percentage of prediction interval; hence, the complementary setup presented seems to have struggled with it.

The fraction of observed inflows bounded within the estimated prediction interval decreases with increased lead-time (Fig. 9b). Reliability score for lead-times up to 17 hours fulfil the

1 requirement of containing 95% of the observations. For lead-times beyond 17 hours, the  
2 reliability declines and reaches 92% at forecasts lead-time of 24 hours.

3 Findings from evaluation of the forecast skill of the complementary setup using deterministic  
4 and probabilistic metrics support each other. The present procedure is able to improve accuracy  
5 of reservoir inflow forecasts and the level of improvement decreases as the forecast lead-time  
6 increases. Deterministic evaluation of performance of the forecast system indicates that the  
7 concept of complementing the conceptual model with a simple error is not always effective. As  
8 discussed earlier, in some occasions the present method can get counterproductive in  
9 forecasting inflows when the forecast lead-time is beyond 20 hours. Similarly, detailed  
10 assessment of the reliability (Table 5) shows that the *CR* of the forecasting system can get below  
11 95% at forecast lead-times less than 17 hours; e.g. at forecast lead-time of 9 hours only 89% of  
12 the observed spring inflows of year 2006/07 are bracketed in the 95% prediction interval. It can  
13 also be noted that for shorter forecast lead-times, the percentage of observations contained in  
14 the prediction bounds exceed 95%. Although a greater proportion of observations falling in the  
15 prediction bound is desirable, a high *CR* at short forecast lead-times might indicate a too wide  
16 band-width. This along a *CR* that declines with increased lead-time might suggest invalidity of  
17 the assumptions behind computation of the bounds (e.g. Smith et al., 2012). The two issues at  
18 stake here are the Gaussian assumption on the basis of which the prediction bounds were  
19 constructed, and the model identification and parameter estimation approach implemented.  
20 However, the results from the PIT uniformity probability test (Fig. 9c-f) confirm that the  
21 distribution of the forecasts from the complementary framework are consistent with the  
22 observed inflow. The PIT plots of the observed inflows almost plot along the diagonal bisector  
23 and remain within the Kolmogorov 5% significance in cases of slightly deviations. From  
24 operational hydrology point of view, we concur with the opinion of Thyer et al. (2009) that the  
25 toughest goodness-of-fit test the complementary framework has to pass is whether the  
26 predictive distribution is consistent with the observed inflow, which the PIT uniform probability  
27 plots evaluate directly. With regard to the model assumptions, we recall from the Kolmogorov-  
28 Smirnov test (section 3.3) that the residuals from the error forecasting model did not honour the  
29 homoscedasticity assumption even after transformation, which might suggest invalidity of the  
30 assumption. According to Schoups and Vrugt (2010), in hydrologic applications residual series  
31 are often assumed to be independent and identically distributed but these assumptions are  
32 usually violated. In the next section, we briefly assess reliability of the model identification and  
33 parameter estimation approach implemented in this study.

### 3.6 On the implemented parameter estimation technique

The parameter (AR model coefficient(s) and transformation parameters) estimation technique we employed (section 2.2.2) follows a pseudo multi-objective optimization approach, which includes minimizing the sum of squares of the residuals and making sure a homoscedastic residual series. We first employed the Least Square (LS) method to estimate the parameters associated to several AR models (of orders 1 to 3). Since the unit of the inflows (the errors as well) in the transformed space depended on the transformation parameters, and the inclusion of the transformation parameters into the calibration problem posed a challenge to identify the optimal among the candidate AR models, we resorted to the dimensionless Kolmogorov-Smirnov (KS) statistic. The KS metric served as a relative quantitative measure to discriminate between candidate models by measuring how close-to-constant the residual variances' are. As a result, the selected AR model is suboptimal in terms of yielding the least discordance between predictions and observations. Putting aside the issue of (in)validity of the Gaussian assumption, we demonstrate that shortcomings of the present LS and KS (LS-KS) model the probabilistic metrics revealed are not unique to the implemented parameter estimation approach. In order to verify this, we setup an AR model estimated the coefficients and transformation parameters by maximizing the Gaussian maximum likelihood (GML).

An AR(2) model was identified with coefficients and transformation parameters:  $\beta = 1.08$ ,  $\lambda = 0.01$ ,  $a_1 = 1.82$  and  $a_2 = -0.82$ . All the deterministic metrics used in this study confirm performance improvement of a slight degree by the GML based model during calibration and validation. This does not come as a surprise because parameters of the LS-LK based model were suboptimal. On the other hand, the KS test revealed that the maximum distance between the sample line and the theoretical line increased to 0.290, which is higher than the statistic the error transformation using parameterization of the LS-KS model (0.10) yielded. To be fair, comparison of the KS statistics associated to the GML and LS-LK transformation parameters might not be appropriate because the LS-KS based AR model was selected for its low KS statistic. Nevertheless, the KS statistic corresponding to the GML based transformation shows a heteroscedasticity of degree higher than the untransformed residuals (0.13). The PIT uniform probability plots assured that the distribution of the predictions matches the observation very well. Comparison of the CR of the GML and LS-LK based models showed a similar proportion of observations contained in the prediction interval. The CR again reveals the same characteristics of high values at short lead-times and the fraction of observations contained in

the prediction bound declines at longer lead-times. This affirms that validity of the Gaussian assumptions stand out as the main issue requiring further investigation in relation to probabilistic assessment.

#### **4 Concluding remarks**

In the present study, the forecasting system comprising additively setup conceptual and simple error model is presented. Parameters of the conceptual model were left unaltered, as are in most operational setups, and the data-driven model was arranged to forecast the corrective measures to be made to outputs of the conceptual models to provide more accurate inflow forecasts into hydropower reservoirs several hours ahead.

Application to the Krinsvatn catchment revealed that the present procedure could effectively improve forecast accuracy over a 24 hours lead-time. This proves that the efficiency of a flow forecasting system can be enhanced by setting up a data-driven model to complement a conceptual model operating in the simulation mode. Furthermore, the current study reveals that analysing characteristics of the residuals from the conceptual model is important and heteroscedastic behaviour should be addressed before identifying and estimating parameters of the error model. Compared to past studies that applied data-driven and conceptual models in a complementary way, the present procedure is successful in providing acceptably accurate forecast for extended lead-times. It also outlines procedure for extracting useful information from the bias, the persistence and the heteroscedasticity the residual series from the conceptual model exhibited, although the assumption that the residuals from the modelling framework to be random failed to hold.

Results also indicate that probabilistic forecasts can be obtained from deterministic models by constructing uncertainty of the complementary setup based on predictive uncertainty of the simple error model. The uncertainty bound seems to satisfy the reliability requirement when evaluated over the entire forecasting period. Its reliability with respect to forecast lead-time also appears satisfactory for lead-times up to 17 hours. Nevertheless, the season wise assessment revealed that the degree of reliability of the forecasts vary from season to season. Given that the error model essentially makes use of the persistence structure in the residuals from the conceptual model, the present procedure seems to be unable to capture transitions in the hydrograph errors from over- to under-estimation (and vice versa). On the one hand, it was

unveiled that the degree of reliability of the forecasts decline with longer lead-times and the deterministic metrics (*RMSE* and *PVE*) confirmed the same.

In order to address these challenges, a future development can be to explore methodologies for taking care of seasonal variability in the structure of the residual series. Updating the error models periodically can be one solution but care must be taken if the selected updating method makes a Gaussian assumption. Another alternative would be to explore more complex stochastic models for the residuals, that use exogenous predictor variables either observed directly (much like the seasonal reservoir inflow forecasting models described in Sharma et al, 2000), or using state variables simulated from the conceptual model (like the Hierarchical Mixtures of Experts framework in Marshall et al, 2006 and Jeremiah et al, 2013). Formulation of these models will also offer better insight into the deficiencies that exist within the HBV conceptual model, thereby allowing further improvement to reduce the structural errors present.

Another interesting topic of future investigation is the intercomparison of the probabilistic forecasts presented in the current paper with the same from popular methods such as Bayesian forecasting system (BFS), the generalized likelihood uncertainty estimation (GLUE) and the Bayesian recursive estimation (BaRE). We believe this would enable identification of the most effective and reliable probabilistic forecasting method that can also be implemented in an operational setup.

## **Acknowledgements**

This work was supported by the Norwegian Research Council through the project Updating Methodology in Operational Runoff Models (192958/S60) and the consortium of Norwegian hydropower companies led by Statkraft. The hydrological data used in the project were retrieved from database of the Norwegian Water Resources and Energy Directorate (NVE). The meteorological data were obtained from Trønderenergi AS and we thank Elena Akhtari for making them available to us. We would like to acknowledge the assistance of Professor Keith Beven in the preparation of this manuscript.

## 1   **References**

- 2   Abebe, A. J. and Price, R. K.: Managing uncertainty in hydrological models using  
3   complementary models, *Hydrolog. Sci. J.*, 48(5), 679-692, 2003.
- 4   Aronica, G. T., Candela, A., Viola, F., and Cannarozz, M.: Influence of rating curve uncertainty  
5   on daily rainfall - runoff model predictions, *Predictions in Ungauged Basins: Promise and*  
6   *Progress*, 303, 116-124, 2006.
- 7   Bergström, S.: The HBV model, in: *Computer Models of Watershed Hydrology*, edited by:  
8   Singh, V.P., Water Resources Publications, Highlands Ranch, CO., 443-476, 1995.
- 9   Beven, K.: *Environmental Modelling: An Uncertain Future?*, Taylor and Francis Group,  
10   London and New York, 2009.
- 11   Beven, K.: *Rainfall-runoff modelling: The primer*, 2nd ed., Wiley-Blackwell, Chichester, 2012.
- 12   Beven, K. J., Smith, P.J., and Freer, J.: So just why would a modeller choose to be incoherent?,  
13   *J. Hydrol.*, 354, 15-32, 2008.
- 14   Box, G.E.P. and Cox, D.R.: An analysis of transformations. *J. Roy. Stat. Soc. B Met.*, 211-252,  
15   1964.
- 16   Chatfield, C.: *Time-series forecasting*, CRC Press, 2000.
- 17   Engeland, K., Xu, C.-Y., and Gottschalk, L.: Assessing uncertainties in a conceptual water  
18   balance model using Bayesian methodology, *Hydrolog. Sci. J.*, 50(1), 45-63, 2005.
- 19   Goswami, M., O'Connor, K. M., Bhattarai, K. P., and Shamseldin, A. Y.: Assessing the  
20   performance of eight real-time updating models and procedures for the Brosna River, *Hydrol.*  
21   *Earth Syst. Sc.*, 9(4), 394-411, 2005.
- 22   Jeremiah, E., Marshall, L., Sisson, S. A., and Sharma, A.: Specifying a hierarchical mixture of  
23   experts for hydrologic modeling: Gating function variable selection, *Water Resour. Res.*, 49(5),  
24   2926-2939, 2013.
- 25   Kachroo, R. K.: River flow forecasting: Part 1 - A discussion of the principles, *J. Hydrol.*, Vol.  
26   133, 1-15, 1992.
- 27   Krzysztofowicz, R.: Bayesian theory of probabilistic forecasting via deterministic hydrologic  
28   model, *Water Resour. Res.*, 35(9), 2739-2750, 1999.



- Krzysztofowicz, R.: The case for probabilistic forecasting in hydrology, *J. Hydrol.*, 249(1–4), 2–9, 2001.
- Li, L., Xu, C. Y., and Engeland, K.: Development and comparison in uncertainty assessment based Bayesian modularization method in hydrological modeling. *J. Hydrol.*, 486, 384–394, 2013.
- Laio, F., and Tamea, S.: Verification tools for probabilistic forecasts of continuous hydrological variables, *Hydrol. Earth Syst. Sci.*, 11(4), 1267 – 1277, 2007.
- Liu, Y., Weerts, A. H., Clark, M., Hendricks Franssen, H.-J., Kumar, S., Moradkhani, H., Seo, D.-J., Schwanenberg, D., Smith, P., van Dijk, A. I. J. M., van Velzen, N., He, M., Lee, H., Noh, S. J., Rakovec, O., and Restrepo, P.: Advancing data assimilation in operational hydrologic forecasting: progresses, challenges, and emerging opportunities, *Hydrol. Earth Syst. Sci.*, 16, 3863–3887, doi:10.5194/hess-16-3863-2012, 2012.
- Madsen, H. and Skotner, C.: Adaptive state updating in real-time flow forecasting—a combined filtering and error forecasting procedure, *J. Hydrol.*, 308, 302–312, 2005.
- Marshall, L., Sharma, A., and Nott, D. J.: Modelling the Catchment via Mixtures: Issues of Model Specification and Validation, *Water Resour. Res.*, 42, W11409, doi:10.1029/2005WR004613, 2006.
- Moll, J. R.: Real time flood forecasting on the River Rhine, in: *Proceedings of the Hamburg Symposium on Scientific Procedures Applied to the Planning, Design and Management of Water Resources Systems*, IAHS Publ. no. 147, 265–272, 1983.
- Montanari, A. and Brath, A.: A stochastic approach for assessing the uncertainty of rainfall-runoff simulations. *Water Resour. Res.*, 42, 40(1), W01106, 2004.
- Nash, J. and Sutcliffe, J.: River flow forecasting through conceptual models part I — A discussion of principles, *J. Hydrol.* 10(3), 282–290, 1970.
- Pappenberger, F., Matgen, P., Beven, K. J., Henry, J. B., Pfister, L., and De Fraipont, P.: Influence of uncertain boundary conditions and model structure on flood inundation predictions, *Adv. Water Resour.*, 29, 1430–1449, 2006.
- Petersen-Overleir, A., Soot, A., and Reitan, T.: Bayesian Rating Curve Inference as a Streamflow Data Quality Assessment Tool, *Water Resour. Manage.*, 23(9), 1835–1842, 2009.

- 1 [Pokhrel, P., Robertson, D.E. and Wang, Q.J.: A Bayesian joint probability post-processor for](#)
- 2 [reducing errors and quantifying uncertainty in monthly streamflow predictions. Hydrol. Earth](#)
- 3 [Syst. Sci. Discuss. 9\(10\), 11199-11225, 2012.](#)
- 4 Renard, B., Kavetski, D., Kuczera, G., Thyer, M., and Franks, S. W.: Understanding predictive
- 5 uncertainty in hydrologic modeling: The challenge of identifying input and structural errors,
- 6 Water Resour. Res., 46, W05521, doi:10.1029/2009WR008328, 2010.
- 7 Roald, L. A., Skaugen, T. E., Beldring, S., Væringstad, Th., Engeset, R., and Førland, E. J.:
- 8 Scenarios of annual and seasonal runoff for Norway based on climate scenarios for 2030-49,
- 9 met.no Report 19/02 KLIMA, 2002.
- 10 [Schoups, G. and Vrugt, J.A.: A formal likelihood function for parameter and predictive](#)
- 11 [inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors.](#)
- 12 [Water Resources Research 46\(10\), 2010.](#)
- 13 Serban, P. and Askew A.J.: Hydrological forecasting and updating procedures, Hydrology for
- 14 the Water Management of Large River Basins, IAHS Publ. n. 201, 357- 369, 1991.
- 15 Shamseldin, A.Y. and O'Connor, K.M.: A non-linear neural network technique for updating of
- 16 river flow forecasts, Hydrol. Earth Syst. Sc., 5(4), 577-597, 2001.
- 17 Sharma, A., Luk, K. C., Cordery, I., and Lall, U.: Seasonal to interannual rainfall probabilistic
- 18 forecasts for improved water supply management: Part 2 - Predictor identification of quarterly
- 19 rainfall using ocean-atmosphere information, J. Hydrol., 239(1-4), 240-248, 2000.
- 20 Shrestha, D.L. and Solomatine, D.P.: Machine learning approaches for estimation of prediction
- 21 interval for the model output. Neural Networks, 19(2), 225-235, 2006.
- 22 Sikorska, A. E., Scheidegger, A., Banasik, K., and Rieckermann, J.: Considering rating curve
- 23 uncertainty in water level predictions, Hydrol. Earth Syst. Sci., 17, 4415–4427, 2013.
- 24 Smith, P. J., Beven, K. J., Weerts, A. H., and Leeda, D.: Adaptive correction of deterministic
- 25 models to produce probabilistic forecasts, Hydrol. Earth Syst. Sci., 16, 2783–2799, 2012.
- 26 Solomatine D. P. and Shrestha D. L.: A novel method to estimate model uncertainty using
- 27 machine Learning techniques. Water Resour. Res., 45, W00B11, 2009.
- 28 [Thyer, M., Renard, B., Kavetski, D., Kuczera, G., Franks, S. W., and Srikanthan, S.: Critical](#)
- 29 [evaluation of parameter consistency and predictive uncertainty in hydrological modeling: a case](#)

1 [study using Bayesian total error analysis, Water Resour. Res., 45, W00B14,](#)  
2 [doi:10.1029/2008WR006825, 2009.](#)

3 Todini, E.: Hydrological catchment modelling: past, present and future, Hydrol. Earth Syst. Sc.,  
4 11(1), 468-482, 2007.

5 Toth, E., Brath, A., and Montanari A.: Real-time flood forecasting via combined use of  
6 conceptual and stochastic models, Physics and Chemistry of the Earth, B, 24(7), 793-798, 1999.

7 [Wang, Q. J., Robertson, D. E., and Chiew, F. H. S.: A Bayesian joint probability modeling](#)  
8 [approach for seasonal forecasting of streamflows at multiple sites, Water Resour. Res., 45,](#)  
9 [W05407, doi:10.1029/2008WR007355, 2009.](#)

10 World Meteorological Organization: Simulated real-time intercomparison of hydrological  
11 models, WMO Pub., 241 pp, 1992.

12 Xiong, L. and O'Connor, K. M.: Comparison of four updating models for real-time river flow  
13 forecasting, Hydrolog. Sci. J., 47(4), 621-639, 2002.

14 Xiong, L. H., Wan, M., Wei, X. J., and O'Connor, K. M.: Indices for assessing the prediction  
15 bounds of hydrological models and application by generalised likelihood uncertainty  
16 estimation. Hydrolog. Sci. J., 54(5), 852-871, 2009.

17 Xu, C-Y.: Statistical analysis of parameters and residuals of a conceptual water balance  
18 model—methodology and case study. Water Resour. Manage., 15, 75–92, 2001.

19

1 Table 1 Model parameters and corresponding optimized values.

Parameter	Description	Unit	Optimized value
<b>Snow routine</b>			
TX	Threshold temperature for rain/snow	[°C]	2.23
CX	Degree-day factor for snow melt (forest free part)	[mm/d°C]	9.95
CXF	Degree-day factor for snow melt (forested part)	[mm/d°C]	5.21
TS	Threshold for snow melt/freeze (forest free part)	[°C]	0.73
TSF	Threshold for snow melt/freeze (forested part)	[°C]	-1.80
CFR	Refreeze coefficient	[mm/d°C]	0.04
LW	Max relative portion liquid water in snow	[-]	0.085
<b>Soil and evaporation routine</b>			
FC	Field capacity	[mm]	306.87
FCDEL	Minimum soil moisture filling for POE	[-]	0.31
BETA	Non-linearity in soil water retention	[-]	3.84
INFMAX	Infiltration capacity	[mm/h]	30.22
<b>Groundwater and response routine</b>			
KUZ2	Outlet coefficient for quickest surface runoff	[1/day]	1.65
KUZ1	Outlet coefficient for quick surface runoff	[1/day]	0.99
KUZ	Outlet coefficient for slow surface runoff	[1/day]	0.42
KLZ	Outlet coefficient for groundwater runoff	[1/day]	0.09
PERC	Constant percolation rate to groundwater storage	[mm/day]	1.60
UZ2	Threshold between quickest and quick surface runoff	[mm]	122.34
UZ1	Threshold between quick and slow surface runoff	[mm]	49.97

1 Table 2 Summary of overall and seasonal performance of the conceptual model during the  
2 calibration (2001/02 to 2005/06) and validation (2006/07 to 2010/11) periods.

Seasons	Calibration period			Validation period		
	RMSE [mm]	RE [%]	NSE [-]	RMSE [mm]	RE [%]	NSE [-]
Overall	0.139	1	0.842	0.162	18.8	0.700
Autumn	0.147	1.8	0.724	0.147	11.3	0.769
Winter	0.182	-3.7	0.894	0.126	9.7	0.812
Spring	0.131	-2.7	0.709	0.246	24.6	0.509
Summer	0.073	28.2	0.641	0.079	38.2	0.592

3

1 Table 3 Ratio between occurrence frequency of low PVE ( $\leq 10\%$ ) and high PVE ( $> 10\%$ ) errors  
2 for the hydrologic years 2006/07-2010/11.

Data set	Overestimation				Underestimation			
	Aut.	Win.	Spr.	Sum.	Aut.	Win.	Spr.	Sum.
Simulated forecast (HBV model)	4.4	5.1	7.6	4.5	6.2	5.2	12.8	25.4
Forecast (complementary setup)	1.1	1.2	1.5	2.0	0.9	0.5	1.1	1.3

3

1 Table 4 Relative RMSE reductions (%) in reservoir inflows forecast as a function of forecast lead-time (\* designates relative RMSE reduction  
2 of <0)

	Season /year	Lead Time [hour]																							
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Autumn	06/07	89.3	79.3	70.1	62.7	56.7	52.3	48.5	45	41.7	38.4	35	31.6	28.2	25.6	23.7	21.7	19.1	16.6	15.3	14.3	13.8	13	11.5	10.0
	07/08	91.6	84.4	78.6	73.5	67.6	62.2	58.0	53.8	50.7	48.0	44.8	41.4	38.8	36.3	33.8	30.7	26.3	19.5	10.9	3.3	*	*	*	*
	08/09	93.9	87.9	81.7	76.7	71.0	65.9	62.1	58.5	54.1	49.2	44	39.4	35.7	32.3	28.8	25.7	23.2	70	18.4	16.7	15.3	14.1	12.7	11.5
	09/10	90.9	83.2	76.9	70.9	64.7	59.1	54.9	51.0	47.2	44.2	41.1	38.1	35.1	30.0	29.5	27.1	25.1	23.3	21.9	70.0	70.0	10.0	19.1	18.4
	10/11	92.1	84.9	78.7	67.7	62.4	57	53.9	51.2	47.5	44.8	42.4	40.3	38	35.8	33.9	30.0	29.4	26.2	23.1	30.0	17.2	14.7	12.7	10.9
Winter	06/07	94.2	87.9	82.2	75.6	60.5	49.3	42.8	36.3	31.3	26.3	21.4	17.5	12.9	9.0	6.7	4.6	2.5	1.3	1.0	0.0	*	*	*	*
	07/08	91	81.9	73.3	66.2	59.9	54.1	49.2	44.8	40	36.1	33.3	30.8	28.1	25.4	23.2	90	19.5	17.5	15.6	15.5	16.5	17.5	18.1	18.4
	08/09	91.7	83.9	77.0	74.0	72.2	68.4	62.2	55.1	49.5	44.4	39.8	36	28.9	22.2	18.2	15.6	13.9	12.8	11.9	11.1	9.9	8.6	7.3	5.8
	09/10	94.9	91.4	87.3	83.5	80.3	78.8	76.7	72.7	65.9	58.1	51.8	46.9	43.4	40.2	37.7	35.5	33.7	32.2	30.9	29.4	27.8	26	24.1	22.2
	10/11	93.9	88.7	83.1	75.9	68.1	64.9	61.4	57.1	52.3	47	41.8	36.9	32.2	28.4	26	24.2	22.6	90	19.4	17.7	16	14.6	13	11.1
Spring	06/07	94.2	88.2	82.4	77	71.7	66.3	61.1	56.4	52.3	48.9	45.8	43.1	40.6	38.3	36	33.9	31.8	30	28.5	27.2	26.2	25.2	24.1	23.2
	07/08	96.6	93.3	89.8	86.2	82.6	79.0	75.6	72.8	70.4	68.4	66.6	64.9	63.1	61.3	59.4	57.6	55.8	54	52.5	51.1	49.7	48.4	47.1	46.0

Summer	08/09	95	90.4	85.8	81.6	77.7	73.7	70.6	67.9	65.7	63.5	61.1	58.7	56.3	54	51.7	49.4	47	44.7	42.4	40.1	37.7	35.3	33.2	31.6
	09/10	93.9	87.7	81.7	76.0	70.6	64.9	59.3	54.4	50.6	47.4	44.8	42.5	40.4	38.5	36.8	35.2	33.9	32.8	30.0	31.3	30.5	29.7	29.0	28.3
	10/11	94.6	88.6	82.2	75.7	69.4	63.4	57.7	52.5	48.7	46.8	44.5	41.7	39.0	36.7	34.6	32.7	31.1	29.8	28.7	27.8	26.8	25.8	24.6	23.7
	06/07	94.8	90	85.7	82.8	80.1	76.3	72.6	69.7	67.4	66.0	65.1	63.7	60.1	58.2	56.3	54.2	51.6	49.6	47.6	44.9	42.2	39.5	36.8	34.4
	07/08	90.7	81.4	73.3	66.3	60.3	55.6	51.4	48.0	45.4	42.6	39.9	39.4	39.1	37.1	34.6	32.8	31.0	29.3	28.4	27.4	26.9	26.2	24.8	23.2
	08/09	97.2	94.4	91.6	89	85.1	78.2	69.2	60.3	52.9	47.1	41.6	36.7	32.5	28.8	25.4	22.7	50.0	18.6	17.1	15.9	14.6	13.3	12.4	11.9
	09/10	92.4	84.8	79.1	76.2	74.2	71.5	68.4	65.2	61.0	57.1	54.3	51.9	50.0	47.7	45.1	43.0	41.1	39.3	37.0	35.8	35.0	34.1	33.2	30.0
	10/11	94.2	88.7	82.9	76.4	69.7	64.4	59.3	54.3	49.8	45.8	42.5	39.8	37.2	35.1	33.1	31.5	30.0	28.6	27.5	27.0	26.5	25.9	25.5	25.0

1  
2



1 Table 5 Summary of seasonal containing ratio (95% prediction interval) during reservoir inflow forecasting (2006/07 to 2010/11)

	Season	Lead Time [hour]																							
		/year	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Autumn	06/07	99.9	99.9	97.8	97.8	97.8	97.8	97.8	97.8	97.8	97.8	96.7	94.5	94.5	93.4	93.4	93.4	93.4	90.1	90.1	91.2	90.1	90.1	89.0	89.0
	07/08	99.9	99.9	98.9	98.9	97.8	97.8	97.8	97.8	97.8	97.8	96.7	94.5	91.2	90.1	90.1	89	87.9	87.9	86.8	85.7	85.7	84.6	83.5	83.5
	08/09	99.9	99.9	99.9	99.9	99.9	98.9	98.9	95.6	95.6	95.6	95.6	95.6	95.6	95.6	95.6	95.6	94.5	93.4	93.4	93.4	92.3	92.3	91.2	90.1
	09/10	99.9	99.9	98.9	97.8	97.8	96.7	96.7	95.6	94.5	93.4	93.4	91.2	92.3	92.3	92.3	92.3	93.4	93.4	92.3	92.3	92.3	91.2	90.1	90.1
	10/11	99.9	99.9	99.9	98.9	98.9	97.8	98.9	98.9	97.8	96.7	95.6	95.6	95.6	95.6	95.6	95.6	95.6	94.5	93.4	93.4	93.4	92.3	92.3	91.2
Winter	06/07	99.9	99.9	99.9	99.9	97.8	96.7	96.7	95.6	95.6	95.6	95.6	95.6	94.4	94.4	93.3	93.3	92.2	92.2	92.2	92.2	91.1	91.1	91.1	90.0
	07/08	99.9	99.9	98.9	97.8	97.8	97.8	97.8	97.8	96.7	96.7	94.5	93.4	93.4	92.3	94.5	94.5	94.5	95.6	96.7	95.6	95.6	95.6	94.5	94.5
	08/09	99.9	99.9	99.9	99.9	98.9	98.9	98.9	97.8	97.8	97.8	97.8	97.8	97.8	95.6	95.6	95.6	95.6	94.4	94.4	94.4	94.4	94.4	95.6	95.6
	09/10	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	98.9	98.9	98.9	98.9	98.9	98.9	98.9	98.9	98.9	98.9	97.8	97.8	97.8	97.8
	10/11	99.9	99.9	99.9	99.9	98.9	96.7	96.7	96.7	96.7	96.7	96.7	96.7	96.7	95.6	95.6	96.7	95.6	95.6	95.6	95.6	95.6	94.4	94.4	94.4
Spring	06/07	99.9	99.9	98.9	98.9	97.8	95.7	94.6	93.5	89.1	89.1	89.1	89.1	90.2	88.0	88.0	88.0	88.0	88.0	87.0	85.9	84.8	84.8	84.8	83.7
	07/08	99.9	99.9	99.9	99.9	99.9	99.9	99.9	98.9	98.9	98.9	98.9	98.9	97.8	97.8	97.8	96.7	95.7	94.6	94.6	94.6	94.6	94.6	94.6	94.6
	08/09	99.9	99.9	98.9	98.9	98.9	98.9	97.8	97.8	97.8	96.7	96.7	96.7	96.7	96.7	96.7	96.7	95.7	95.7	95.7	93.5	93.5	93.5	93.5	92.4
	09/10	99.9	99.9	98.9	97.8	97.8	97.8	96.7	96.7	94.6	94.6	94.6	93.5	93.5	93.5	91.3	91.3	91.3	91.3	90.2	90.2	91.3	89.1	89.1	90.2

Summer	10/11	99.9	98.9	98.9	96.7	96.7	95.7	94.6	93.5	92.4	92.4	90.2	90.2	89.1	88	89.1	87	85.9	85.9	84.8	83.7	83.7	83.7	82.6	82.6
	06/07	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	98.9	98.9	98.9	98.9	98.9	98.9	97.8	97.8	97.8	97.8	97.8
	07/08	99.9	99.9	99.9	99.9	98.9	98.9	98.9	98.9	98.9	98.9	98.9	98.9	98.9	98.9	98.9	98.9	98.9	98.9	98.9	98.9	98.9	98.9	98.9	98.9
	08/09	99.9	99.9	99.9	99.9	99.9	99.9	99.9	98.9	98.9	98.9	98.9	98.9	98.9	98.9	98.9	98.9	98.9	98.9	98.9	98.9	98.9	98.9	98.9	98.9
	09/10	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	98.9	98.9	98.9	98.9	98.9
	10/11	99.9	99.9	99.9	99.9	98.9	98.9	98.9	98.9	98.9	97.8	96.7	96.7	96.7	96.7	96.7	96.7	96.7	96.7	96.7	95.7	95.7	95.7	95.7	95.7

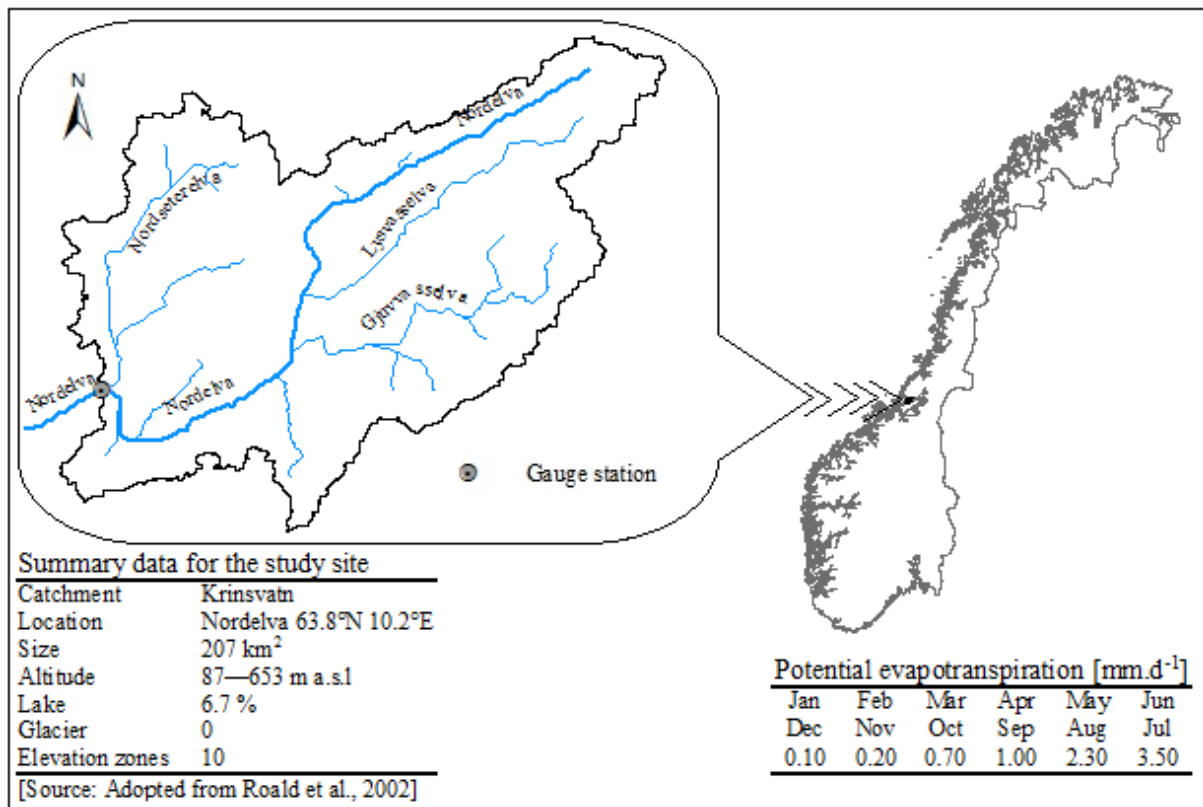
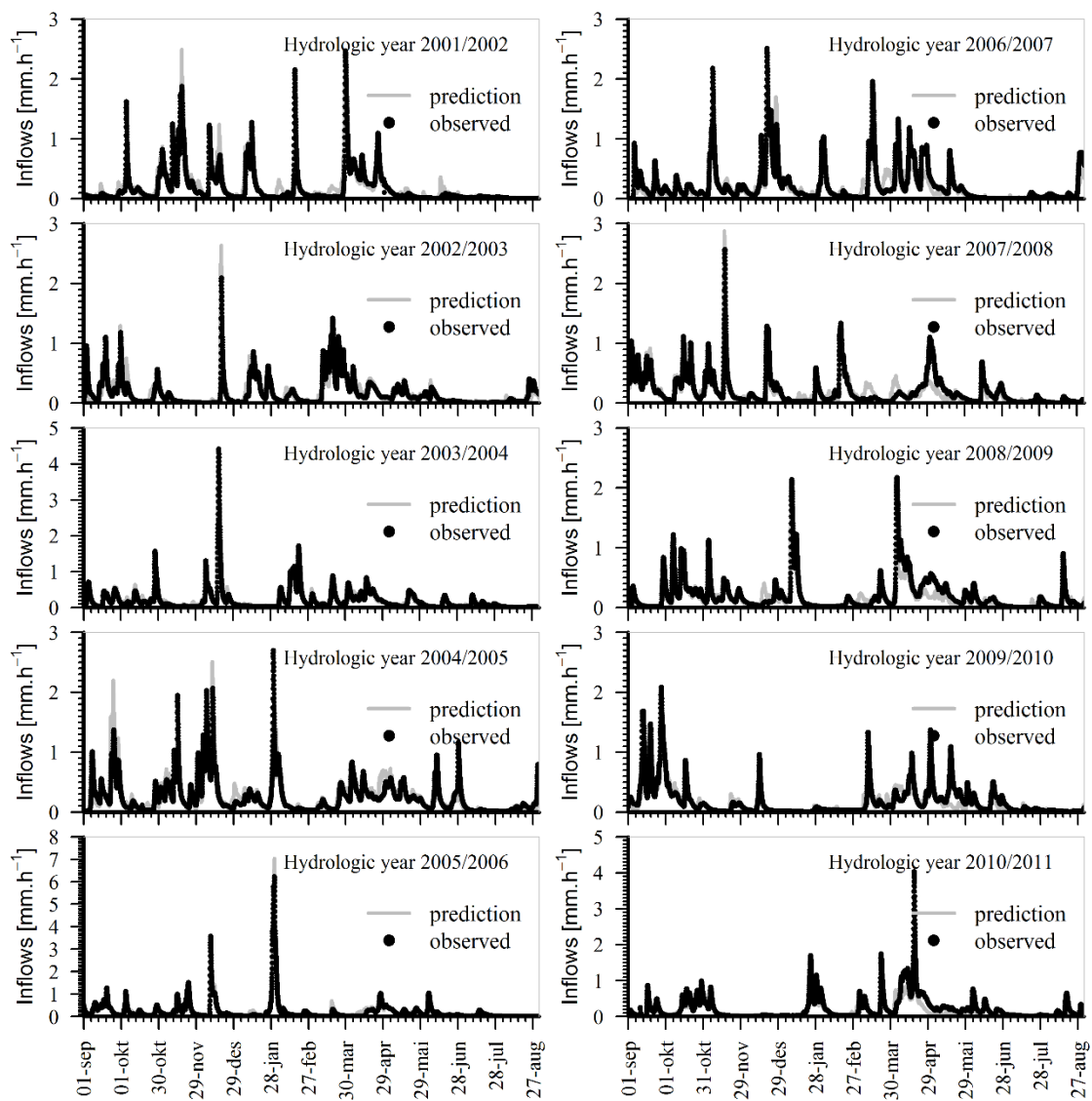


Figure 2. Location, characteristics and potential evapotranspiration estimates of the study catchment.



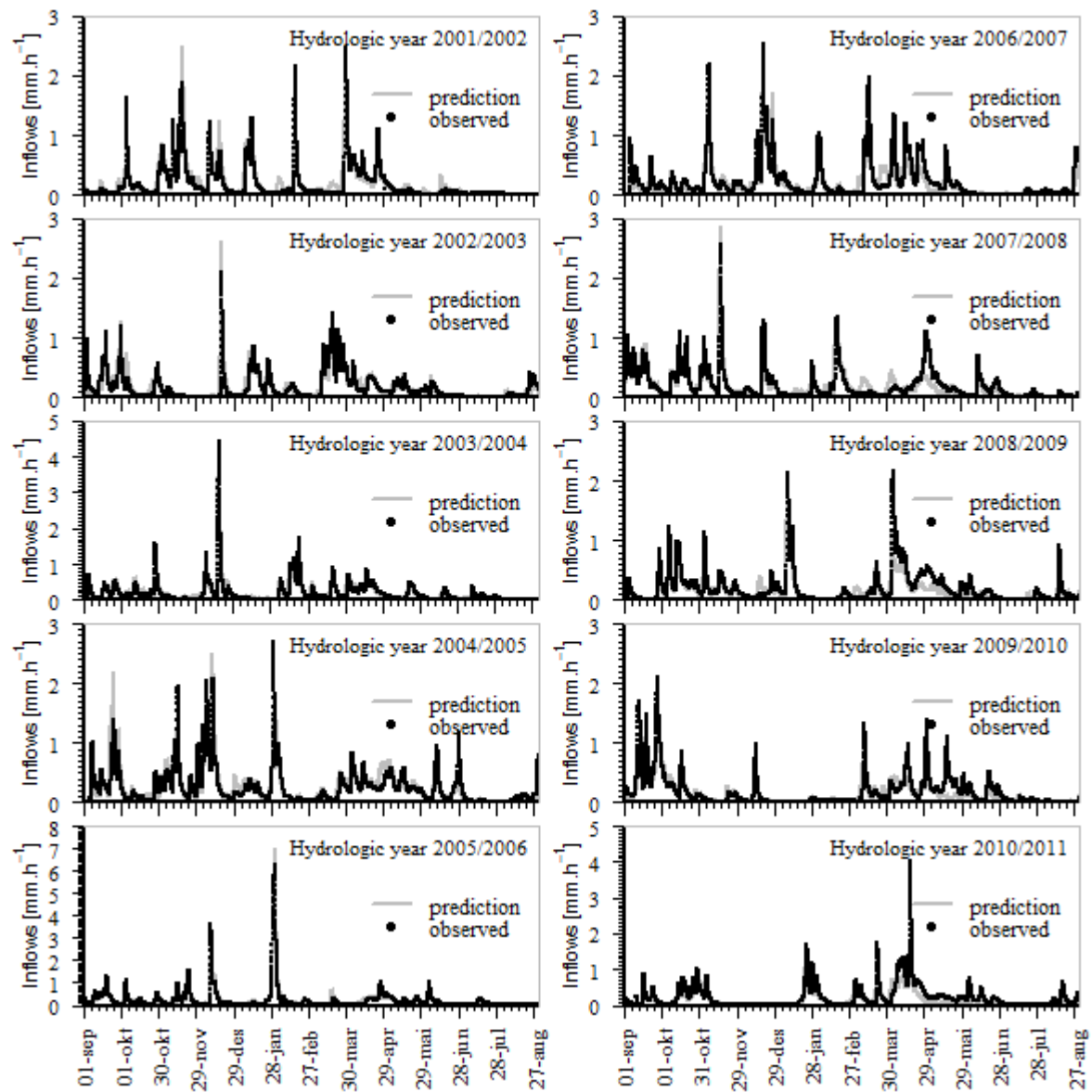


Figure 3. Observed and predicted reservoir inflow hydrographs during calibration (left column) and validation (right column) of the conceptual model.

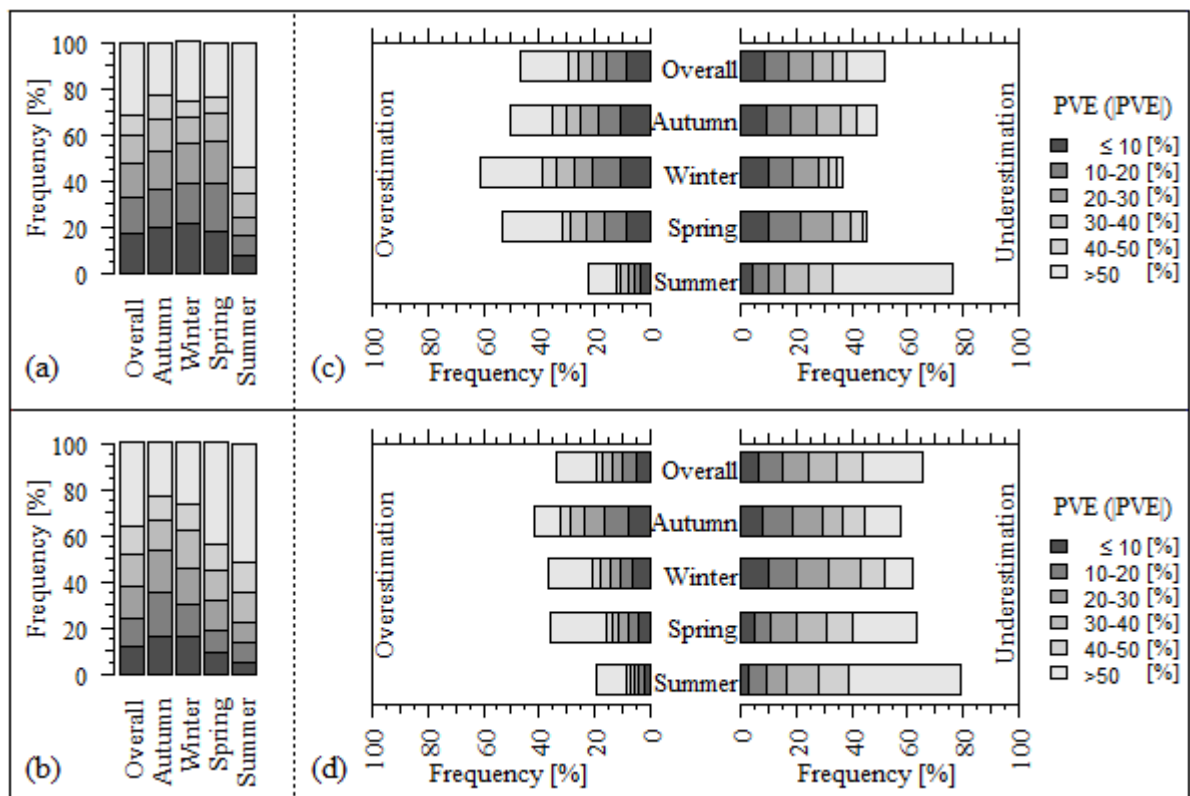
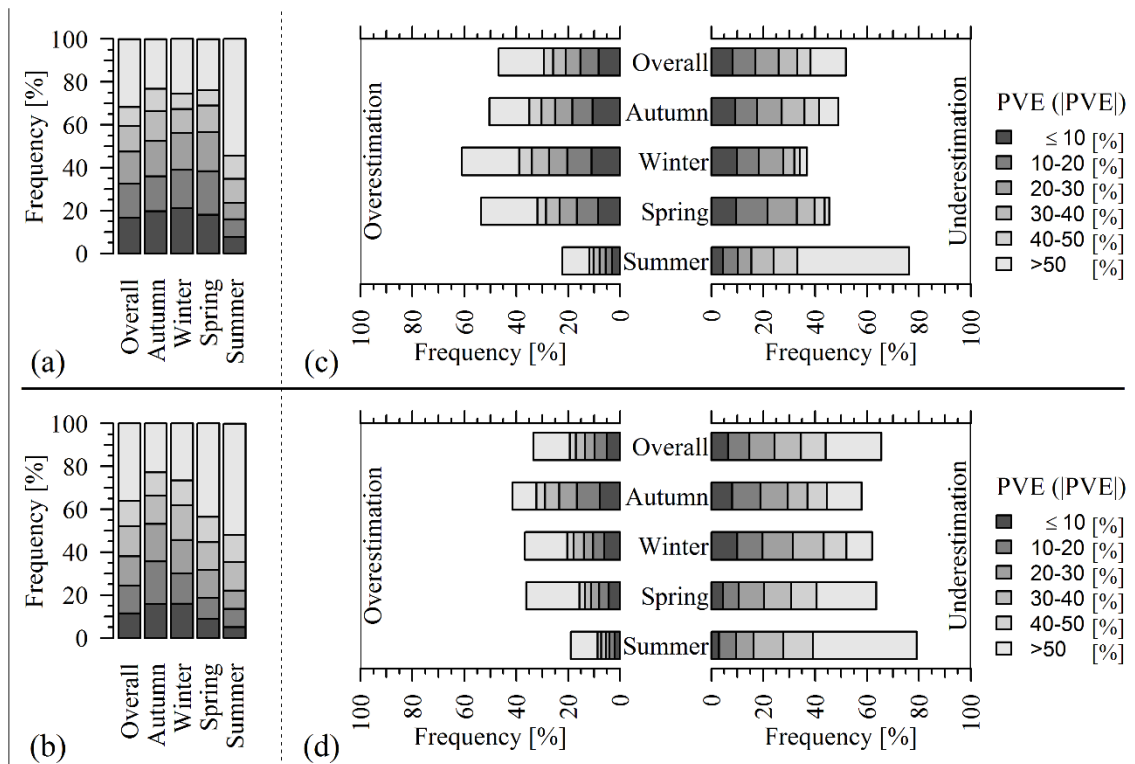
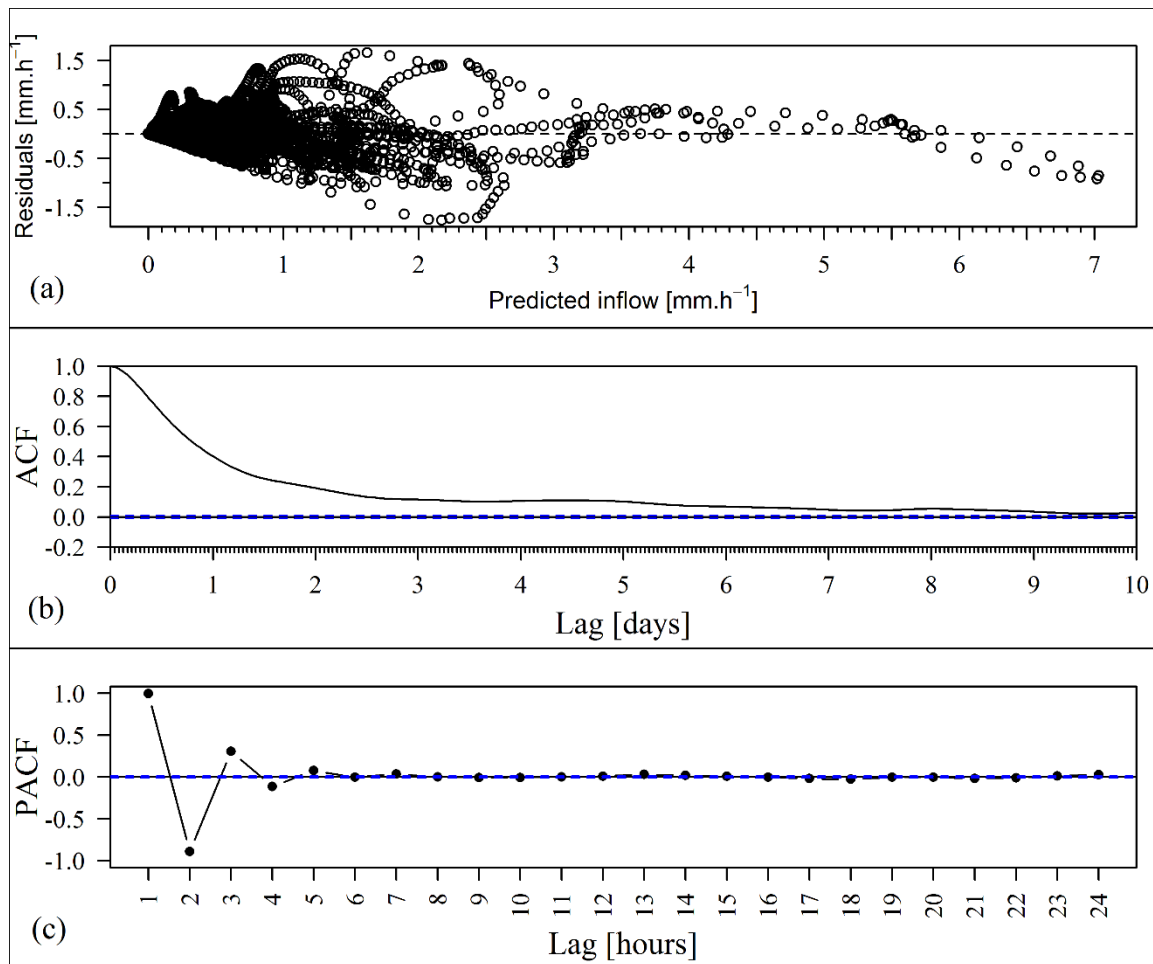


Figure 4. Stacked-column plots of: (1) PVE counts of the six absolute PVE classes ( $\leq 10\%$ , 10-20%, 20-30%, 30-40%, 40-50% and  $>50\%$ ) during calibration (a) and validation (b); and (2)

- 1 the fraction of times under- and over-estimation incidents corresponding to the six PVE classes
- 2 occurred during calibration (c) and validation (d).



1



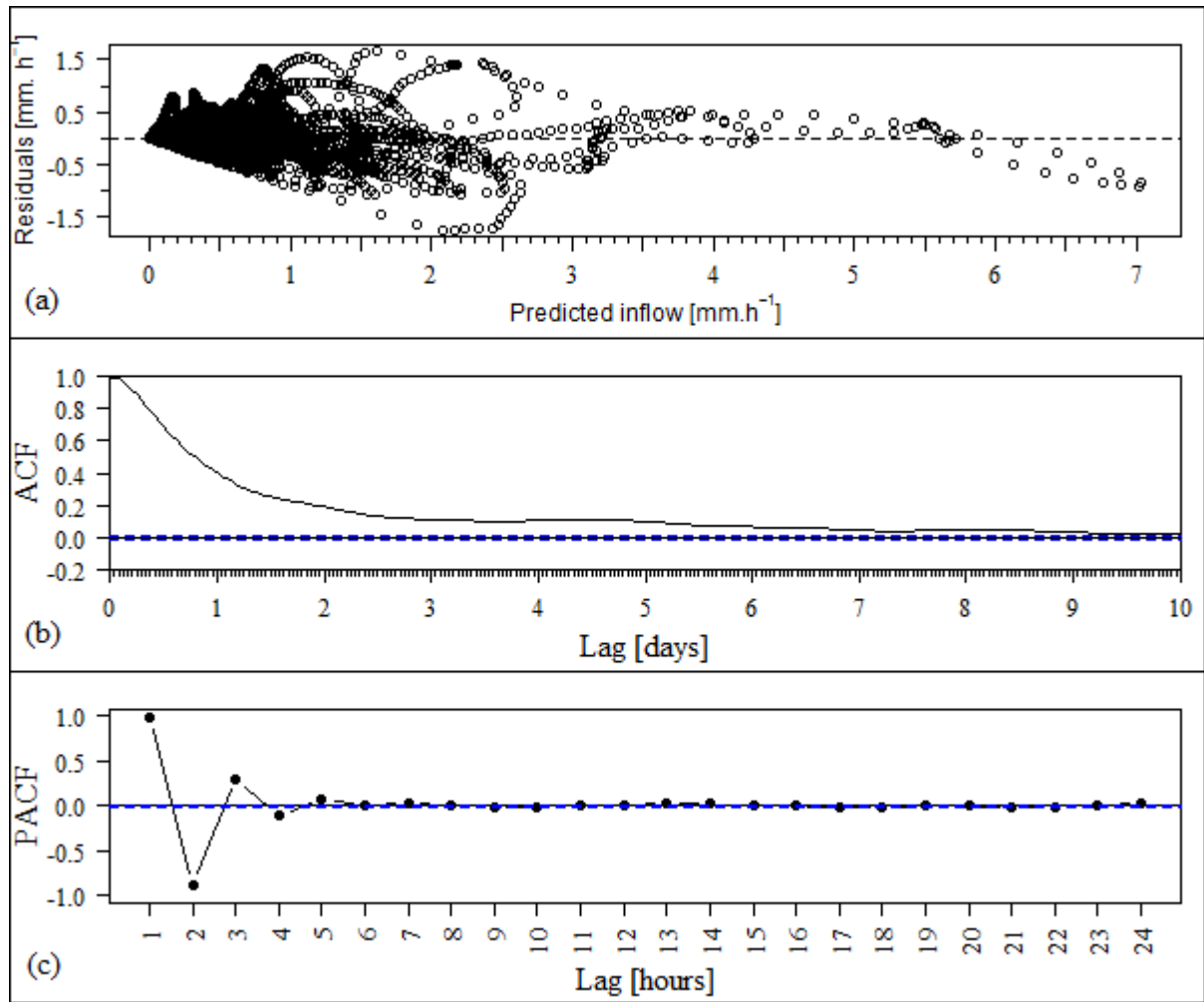


Figure 5. Plots of (a) residuals from the conceptual model as a function of predicted inflow during the calibration period, (b) autocorrelation function of the residuals, and (c) partial autocorrelation functions of the residuals.

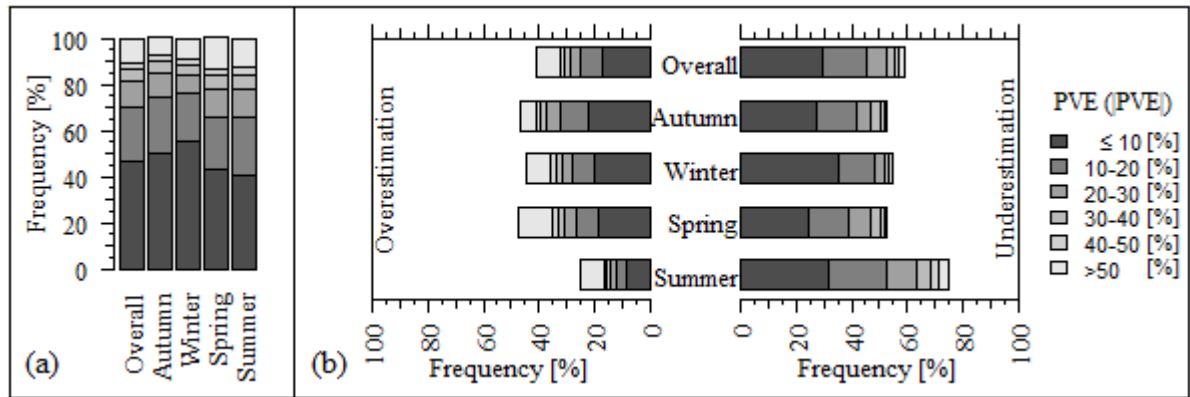
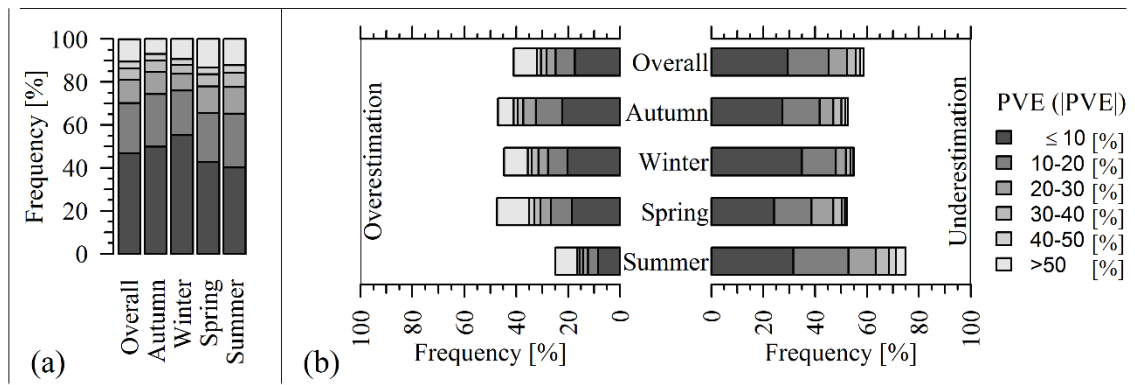
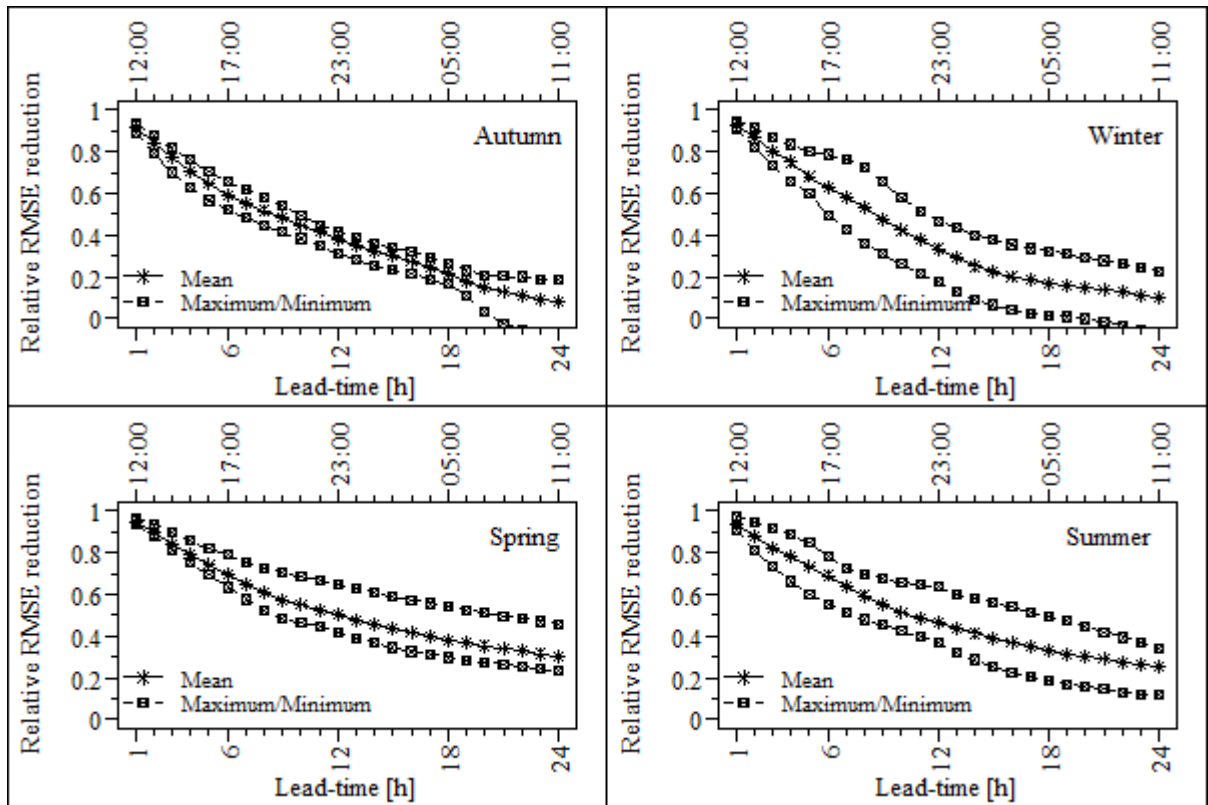
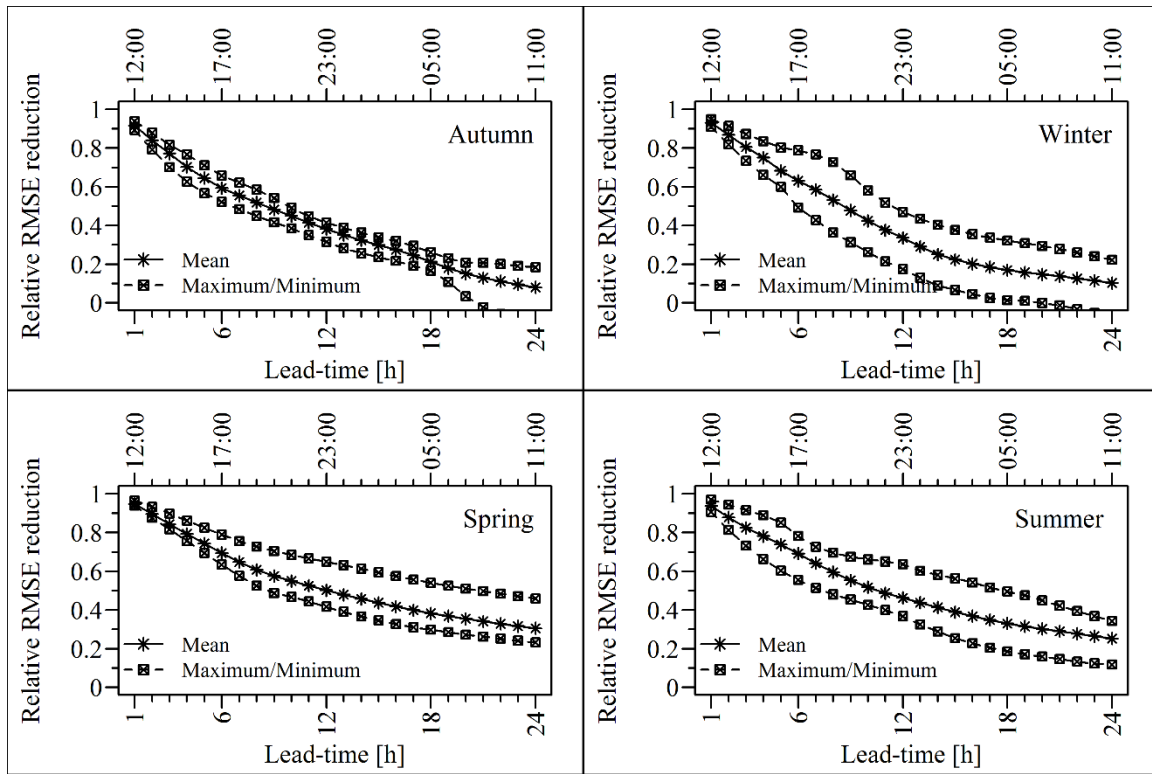
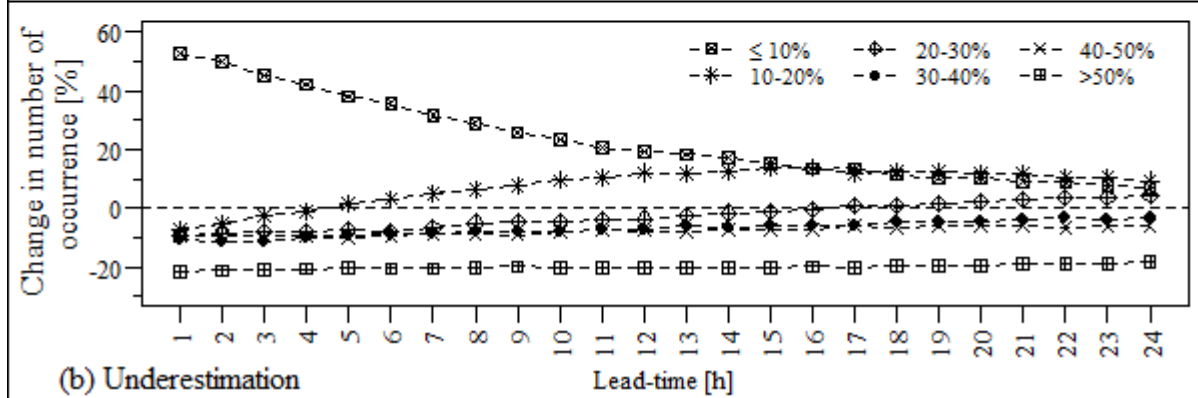
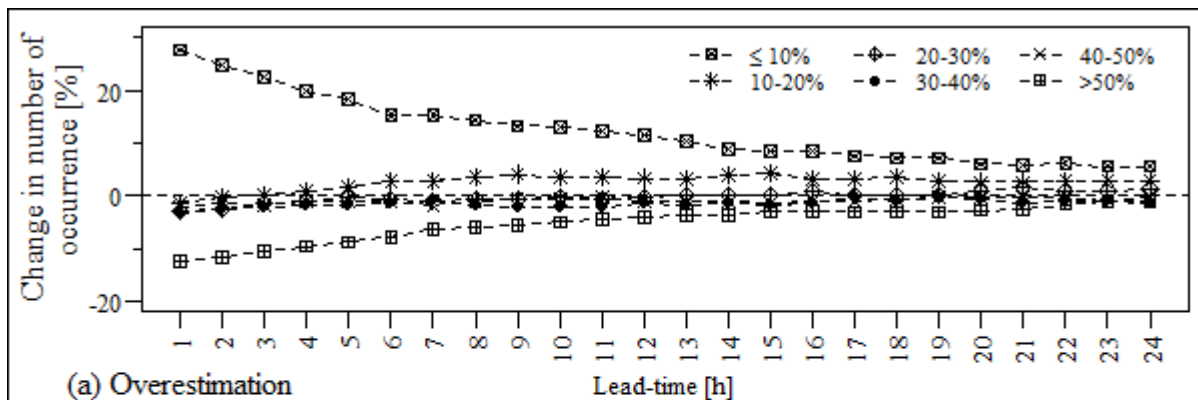
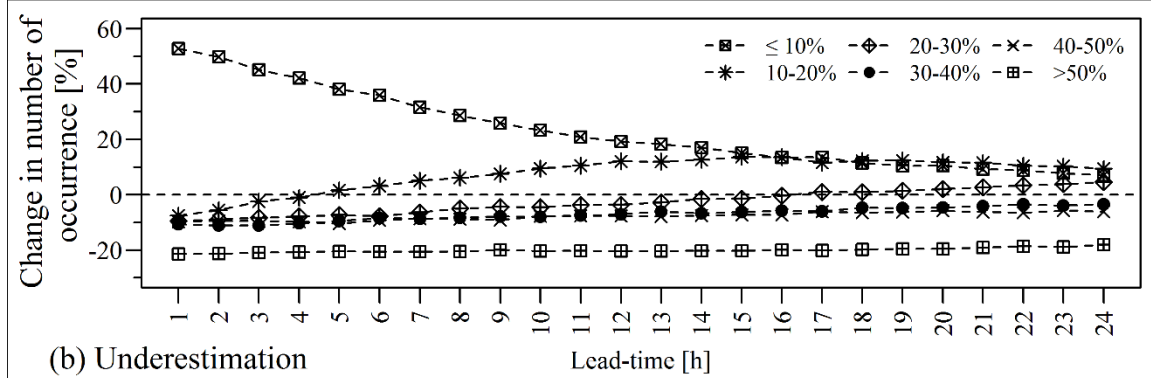
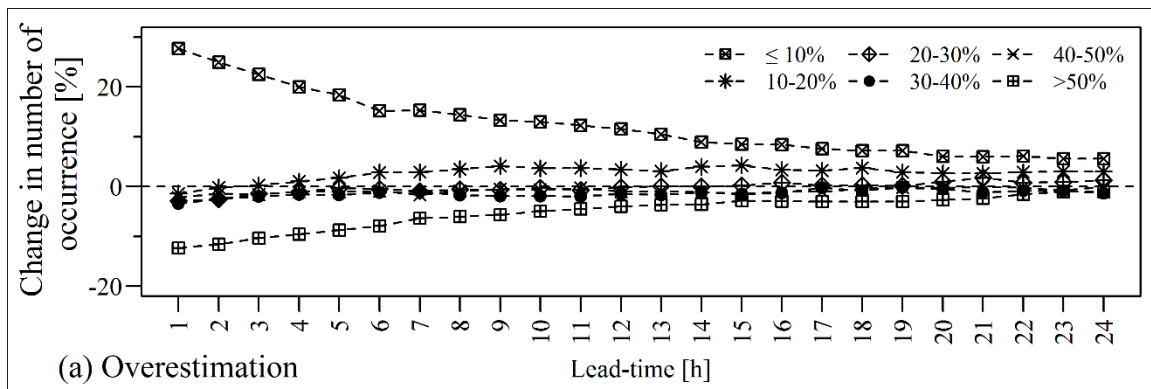


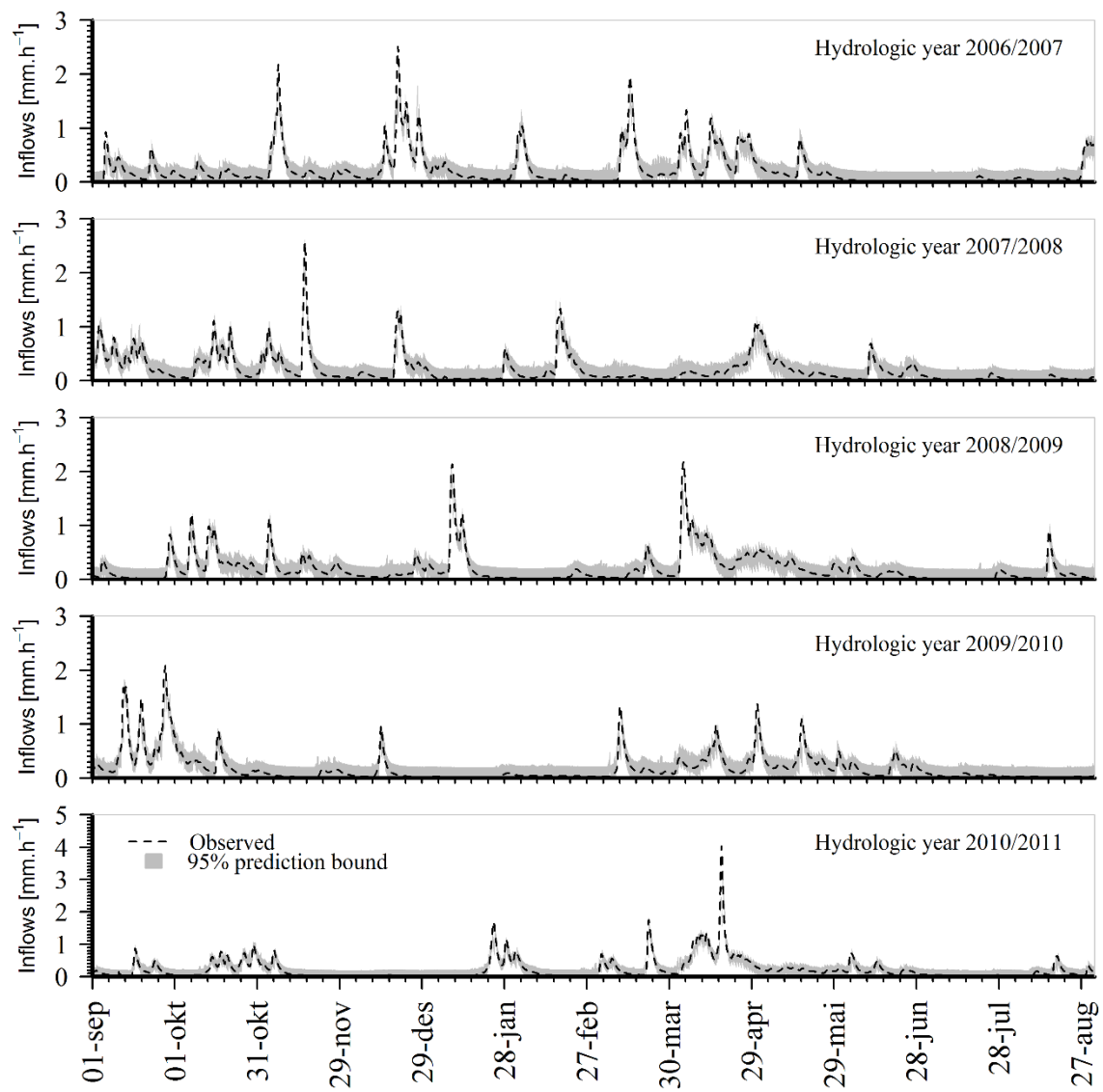
Figure 6. Stacked-column plots of: (a) PVE counts of the six absolute PVE classes ( $\leq 10\%$ , 10-20%, 20-30%, 30-40%, 40-50% and  $>50\%$ ) observed in reservoir inflow forecasts from the complementary setup; and (b) the corresponding fraction of times under- and over-estimation incidents corresponding to the six PVE classes occurred. Hydrologic years 2006/07-2010/11.



1 Figure 7. Summary of relative seasonal RMSE reductions as a function of forecast lead-time  
2 (minimum, mean and maximum values computed from corresponding computations for  
3 hydrologic years 2006/07 - 2010/11).



- 1 Figure 8. Change in number of occurrence of the six absolute PVE classes ( $\leq 10\%$ , 10-20%, 20-
- 2 30%, 30-40%, 40-50% and  $>50\%$ ) as a function of forecast lead-time: (a) overestimation and
- 3 (b) underestimation.



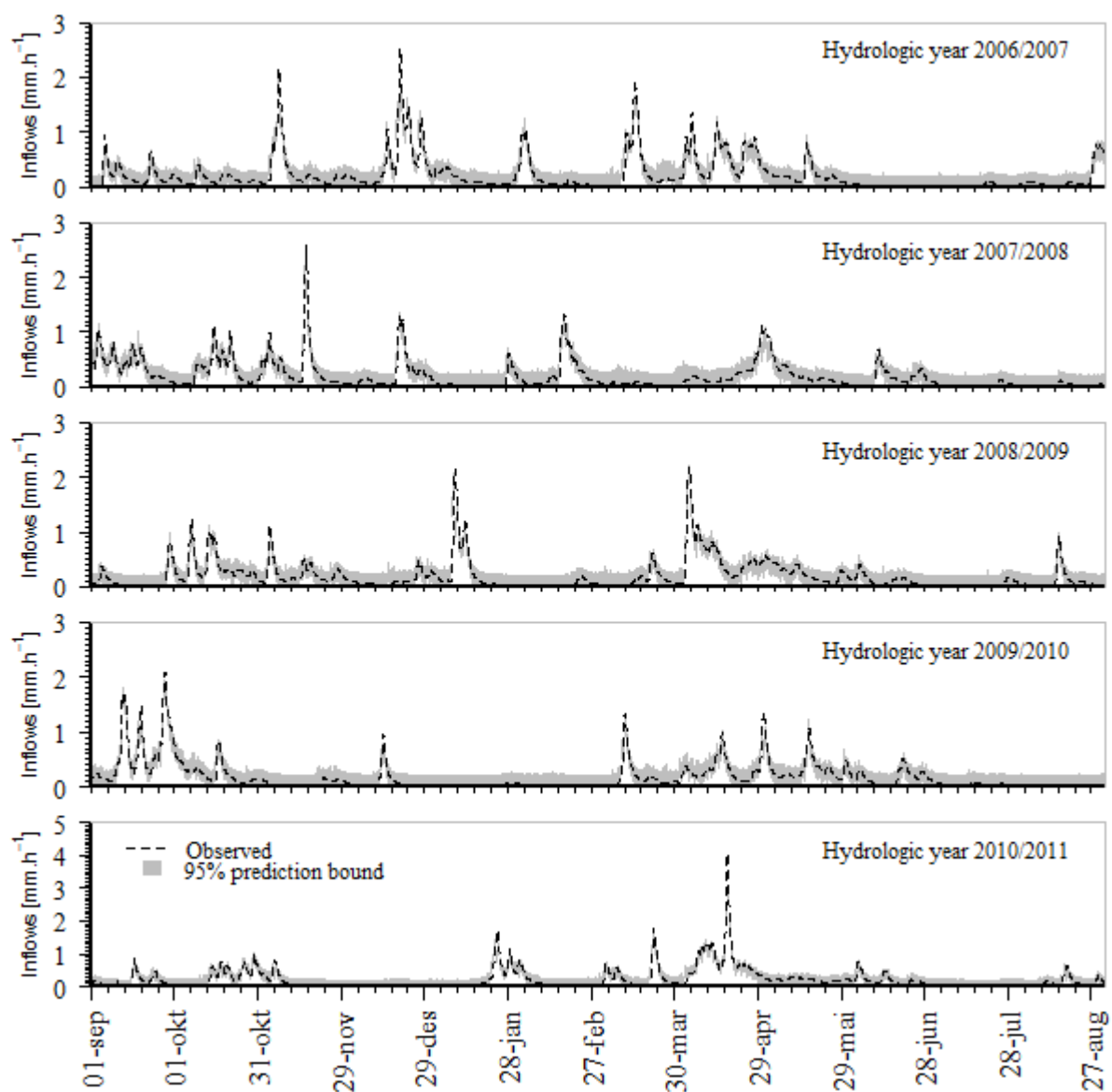
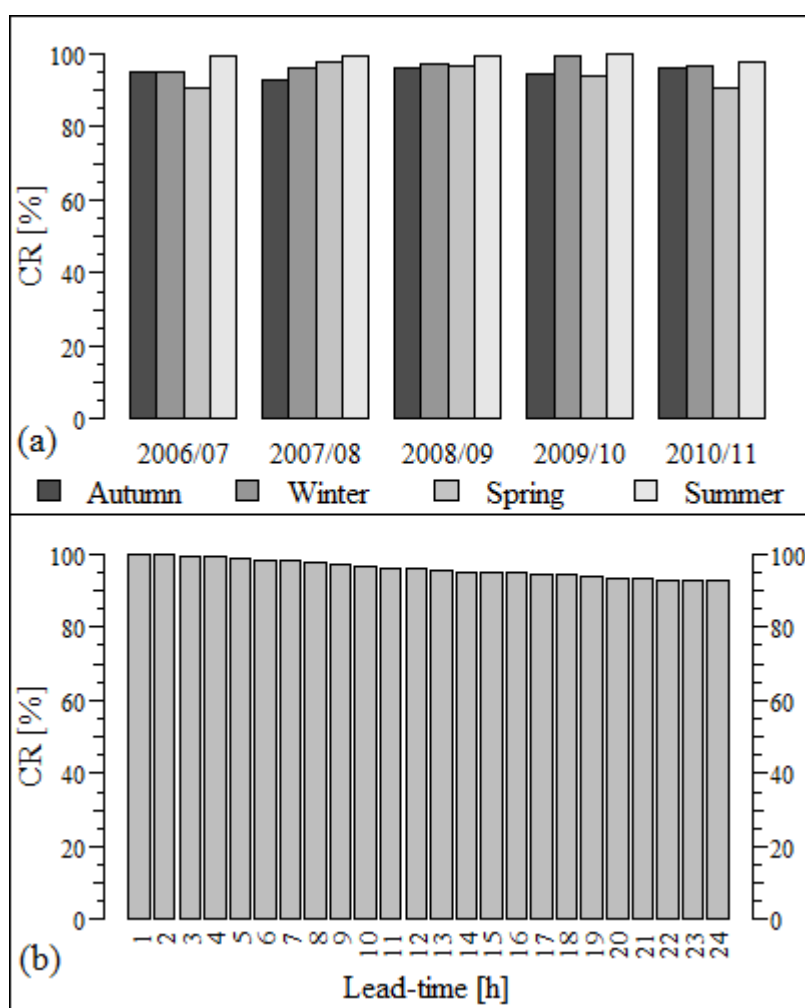


Figure 9. Observed hydrograph (broken lines) and the 95% prediction bound





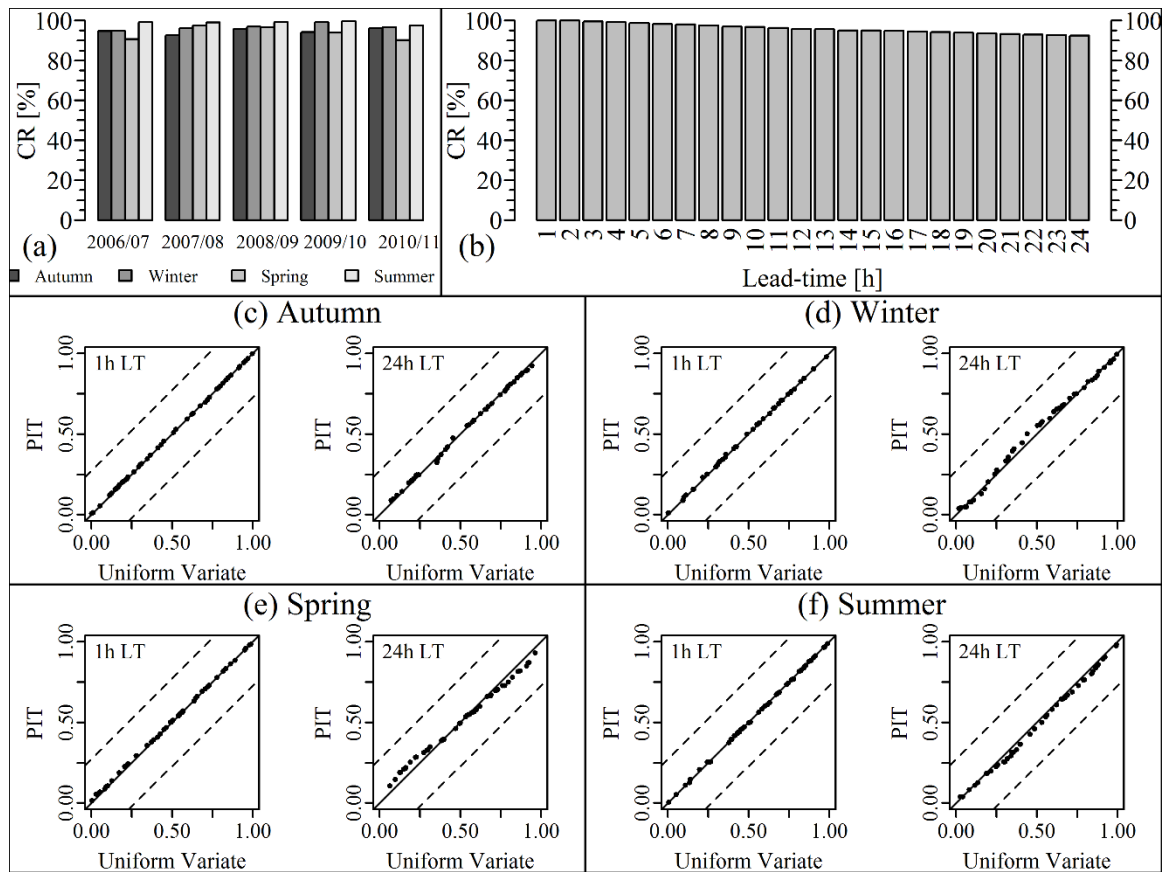


Figure 10. Reliability score (containing ratio-CR) for 95% prediction interval for: (a) each season of every hydrologic year; and (b) different forecast lead-times based on entire series. Panels (c)-(f): sample PIT uniform probability plots for each of the four seasons at 1 hour and 24 hour forecast lead-times. Solid line designates the theoretical uniform distribution, broken lines represent the Kolmogorov significance band, and the dots denote PIT value of the observed inflow.