ISSN 0097-8078, Water Resources, 2018, Vol. 45, Suppl. 1, pp. S90–S101. © Pleiades Publishing, Ltd., 2018.

# Validation of a Hydrological Model Intended for Impact Study: Problem Statement and Solution Example for Selenga River Basin

**A. N. Gelfan[a, b] and T. D. Millionshchikova[a, *]**

[a] *Water Problems Institute, Russian Academy of Sciences, Moscow, 119333 Russia*

[b] *Department of Land Hydrology, Faculty of Geography, Moscow State University, Moscow, 119991 Russia*

**e-mail: tatyana.million@mail.ru*

Received June 19, 2018

**Abstract**—The study is aimed to evaluate a hydrological simulation model intended for assessing climate change impact. A new test was suggested and applied to evaluate the performance of a physically based model of Selenga River runoff generation. In this test, to calibrate the model, an enhanced Nash–and-Sutcliffe efficiency (NSE) criterion was used, including trend-oriented reference (benchmark) models instead of the simple reference model used in the original NSE criterion. Next, modifications were made in the Differential Split Sample test (DSS-test) of V. Klemeš (1986), focused on differences in the model performance criteria for climatically contrasting periods, and a new statistical measure was proposed to estimate the significance of these differences. After that, model performance was evaluated for four sites within the catchment, three indicators of interest (daily, monthly, and annual discharge series), and the model ability to reproduce the observed trends in annual and seasonal discharge values was assessed. The model proved robust enough to be applied to assessing climate change impact on the annual and monthly runoff in different parts of the Selenga River basin.

**Keywords:** hydrological modeling, Selenga River Basin, validation, model robustness, impact assessment, crash-test

**DOI:** 10.1134/S0097807818050354

## INTRODUCTION

Predicting the hydrological response to changes in the external impact (for instance, because of climate change) and/or in internal parameters (for instance, because of a change in watershed characteristics) is a traditional problem of terrestrial hydrology. The opportunities to solve these problems with the help of hydrological models developed for unchanged conditions formed the core of major scientific initiatives of The International Association of Hydrological Sciences (IAHS) in the recent two decades. While the previous IAHS decade (2003–2012: Prediction in Ungauged Basins, PUB) was devoted to analyzing the opportunities of "transferring" a hydrological model over space—from gauged basins to ungauged ones, the current IAHS decade (2013–2022: Panta Rhei) is aimed at searching for methods of "transferring" the model over time—into different climate conditions. In both cases, hydrologists have to judge the model performance in the absence of observation data. This paper is devoted to the problem of evaluating a hydrological simulation model intended for assessing climate change impact; more specifically, we will try to answer the question: how should model evaluation be done in the context of climate impact assessment? We emphasize here the difference in the opportunities of evaluation between simulation and forecasting models: the latter can be verified with the use of the available observation data, whereas a simulation model usually cannot be tested directly against some data (the latter, most probably, never be available).

In the following section, before proceeding to the motivation of the paper, its purposes, and the methods used, we concretize the problem statement.

## VALIDATION OF A MODEL INTENDED FOR IMPACT ASSESSMENT: PRAGMATIC APPROACH AND "CRASH-TEST" NECESSITIES

*"All Models are Wrong, but Some are Useful"*
*George Box*

The validation of a numerical model of an open natural system, considered as evidence of the adequacy of a model for the given system, is impossible. The roots of this statement lie in the widespread direction of contemporary philosophy—the critical rationalism of Karl Popper. In the field of geophysical sciences, the grounds of this statement were first studied in detail in [21], including the problem of "non-

uniqueness" of a numerical model, which consists in that models with different structures, parameters, and schemes of numerical implementation produce similar test results; the problem of subjectivity of validation criteria, leading to the lack of objective reasons for considering one model superior to another; the so-called problem of "spatial and temporal divergence," appearing due to the need to use a model beyond the limits of existing observations, etc. The idea that the validation of numerical models of natural systems is utterly impossible gained ground in natural sciences (see, e.g., [32]). However, hydrologists, as well as experts in other fields of geophysics, solve practical problems, and, when validating models, are guided by pragmatic considerations, aphoristically generalized by the outstanding British statistician G. Box (see above). Leaving detail aside, the pragmatic approach can be narrowed down to the following statements.

(1) The validation of a hydrological model is impossible in the universal sense, but there is an opportunity to evaluate the suitability of the model for solving the task for which the model was developed (see, for example, [20, 33]). Correspondingly, the term validation is more often replaced by the term evaluation.

(2) The indicated opportunity is realized by means of special tests, which imitate, with the help of the existing observation data, the "target" conditions of model application (for instance, the conditions of climate change), and use the performance criteria, taking into consideration the specificity of the task (see, for example, [1, 5, 38]).

What is meant by special tests and specific performance criteria will be clarified below. It is important to emphasize here that the above statements link the credibility of a model intended to assess climate change impact with the results of the model's evaluation against available (historical) observations. This linkage, which seems quite natural to hydrologists, has recently provoked some lively discussions (see the review in [22]). The accuracy of calculation of the observed runoff hydrographs with the help of calibrated and verified regional hydrological models is often not regarded as an argument in favor of their applicability for impact assessment (see, for example, [2, 36]), and is ironically compared to the ability to "testing how nicely a mathematical marionette can dance to a tune it has already heard" [19]).

Within the pragmatic approach, on the contrary, the quality of reproducing the observation data by a model is an indicator of its credibility. If a model does not "pass" the developed tests with the indicated quality criteria, one can suggest that it can be inappropriate for the "target" conditions of its application. However, successfully "passing" the tests on the existing observation data, although not being in itself an evidence of model suitability, still can be seen as a confirmation that the model is a possible "candidate" for impact assessment, that is, as a necessary but not sufficient condition of the model's applicability. Thus, the pragmatic approach allows formulating the rules that restrict the opportunities of applying a model to solving tasks, for which it has not been intended.

The model focus on solving extrapolation tasks dictates the necessity of using special tests and performance criteria, which should put the model in 'crash' conditions which reveal the limits of its applicability. However, most of the existing performance evaluation methods are based on easy-to-pass tests (such as the Split-Sample (SS) test) on the basis of single-variable observations (such as a hydrograph at the basin outlet) and 'soft' performance criteria (such as typical Nash–Sutcliffe efficiency criterion). As a result, a lot of "good" models have successfully passed such "soft" tests, and this creates an illusion of these models' applicability to impact assessment. A specific test ("crash-test") is required, allowing one to reinforce the model's applicability, distinguish between the models appropriate for impact studies and unsuitable ones, and understand the grounds for the credibility of a given hydrological model.

A system of tests, which imitate the 'target' conditions of the model application, was first proposed by Klemeš (1986). The author proposed the DSS-test (Differential Split-Sample test) for a model intended for impact study. Various modifications of the DSS-test, as well as other procedures for testing hydrological models, are proposed in [7, 8, 14, 37, 38]. When selecting the hydrological variables (indicators) for which the model should be tested, it is necessary to proceed from the specific nature of the target task ("user-oriented approach" in the terminology of V. Klemeš [20]). The ways for constructing a 'crash test' for models intended for hydrological impact assessment were recently considered in [22]. The authors proposed a step-by-step testing procedure based on the ideas of the above pragmatic approach, including the use of contrasting climate data, specific indicators, etc. for hydrological model evaluation.

There are about 20 performance criteria used to evaluate hydrological model performance [17, 38, 41, 42]. However, despite the existence of a large number of criteria and the availability of recommendations for their choice, in most hydrological publications, the authors use only one criterion, regardless of the impact study target, namely, the Nash–Sutcliffe efficiency (NSE) criterion. The disadvantages of this criterion are well known (see, for example, the review in [16]). In particular, they include the low sensitivity to small and medium discharge values, underestimation of the simulated discharge variance, and overestimation of model efficiency. The last drawback is related to the use of a simplified reference (benchmark) model for performance assessment, namely, mean annual observed discharge. The comparison of a

hydrological model with such a simple reference model results in overestimation of model efficiency.

The opportunity to 'enhance' the standard Nash–Sutcliffe criterion relates to the use of a more complex reference model. A natural step towards the complication is a reference model describing the intra-annual or inter-annual climatic variability of runoff characteristics. The first of the two reference models (in the form of a seasonal climatic hydrograph) was proposed in [12], and then recommended by the WMO for evaluation of hydrological model performance [40]. In a number of papers [34, 35], it is shown that such reference model sharply raises the requirements to the hydrological model under test. However, several decades after the publication of the above-mentioned papers [12, 40], hydrological modellers still prefer to use the 'weakest' reference model.

The studies and examples mentioned above serve as the background, and the knowledge gaps, still existing, drive the main motivation for this study. The objective of the paper is to contribute to the model evaluation studies with the focus on the first use of the crash-test procedure [22] and its modification to physically based hydrological model developed for climate change impact assessment in the Selenga River basin—the main tributary of Lake Baikal. The case study was chosen due to the climatically caused changes in the water regime during the last 25–30 years which have led to a decrease in the average inflow to Lake Baikal by 35%. In this regard, it becomes relevant to construct a physically based hydrological model, which allows not only describing the mechanisms of the current changes in the water regime, but can also be used to assess the hydrological consequences of possible climate change in the 21st century.

The remaining part of the paper is organized as follows. A case study is described in the next section. The methodology, including the hydrological model, the data sets used, and the evaluation procedure, is described in Section 4. Results and discussion are presented in Section 5. The overall conclusions and recommendations are given in Section 6.

## CASE STUDY BASIN

The Selenga River is the major tributary of Lake Baikal, contributing 50 to 60% of its surface water inflow [4, 39]; the river is 1024 km long. It is worth noting that the Selenga River Delta is the world's largest continental delta (approximately 600 km$^2$) [3, 24]. The total basin area of the Selenga River is 447 000 km$^2$ (Fig. 1). Its main tributaries are the Uda, Khilok, Dzhida, Temnik, Chikoy, Orkhon, Tuul, and Kharaa rivers.

Runoff formation conditions in the Selenga River basin are very diverse. The upstream (southern) part of the basin is located in Mongolia (approximately 64% of the basin) and covered by extensive grassland steppe. The downstream (northern) part is located in Southern Siberia, Russia, and is mainly covered by forests on permafrost, which is an important source of soil water in summer. Furthermore, a huge part of the basin has mountainous topography, showing a wide elevation range (from 600 to 3000 m).

The local climate is continental. The winters are long, dry, and cold. Mean monthly temperature of January is –23.5°C. The summers are short and relatively warm. The mean monthly temperature of July is 16°C. The annual precipitation amount ranges between 300 and 400 mm over the territory. Most of precipitation falls in July and August (about 45% of the annual precipitation). Rainfall is the major source of Selenga River runoff [11].

River water regime shows a winter low-flow period from November to March, a relatively low spring snowmelt flood, and high rain floods in summer and autumn. For the period of 1980–2013, mean annual discharge at the basin outlet (Kabansk gauging station) is 958 m$^3$/s, and the mean maximum discharge is 2942 m$^3$/s. Since mid-1990s, a many-year low-flow period has been observed, which is recognized to be the longest such period in the historical observations [11].

## METHOD

### *Hydrological Model*

The ECOMAG (ECOlogical Model for Applied Geophysics) is a semi-distributed process-based hydrological model, describing snow accumulation and melting; soil freezing and thawing; water infiltration into unfrozen and frozen soil; evapotranspiration; thermal and water regime of soil, overland, subsurface, and channel flow with a daily time step [28]. The model accounts for measurable watershed characteristics, such as surface elevation, slope, aspect, land cover and land use, soil and vegetation properties. The parameters are spatially distributed by partitioning the watershed into sub-basins (elementary basins). The parameterization of the sub-grid processes is described in [27].

The model was used earlier for hydrological simulations in many river basins with widely varying sizes and characteristics—from small-to-middle size European basins [14, 15, 28] to the large Volga, Lena, Mackenzie, Amur basins with watershed areas exceeding a million km$^2$ [13, 18, 27].

A DEM with 1 × 1 km spatial resolution was used for basin discretization and river network simulation. A total of 675 elementary basins were delineated, with the average area of 670 km$^2$. Model forcing data for each elementary basin were interpolated from 74 weather stations data with the use of inverse-distance method. Most parameters are physically meaningful and have been derived from the available mea-

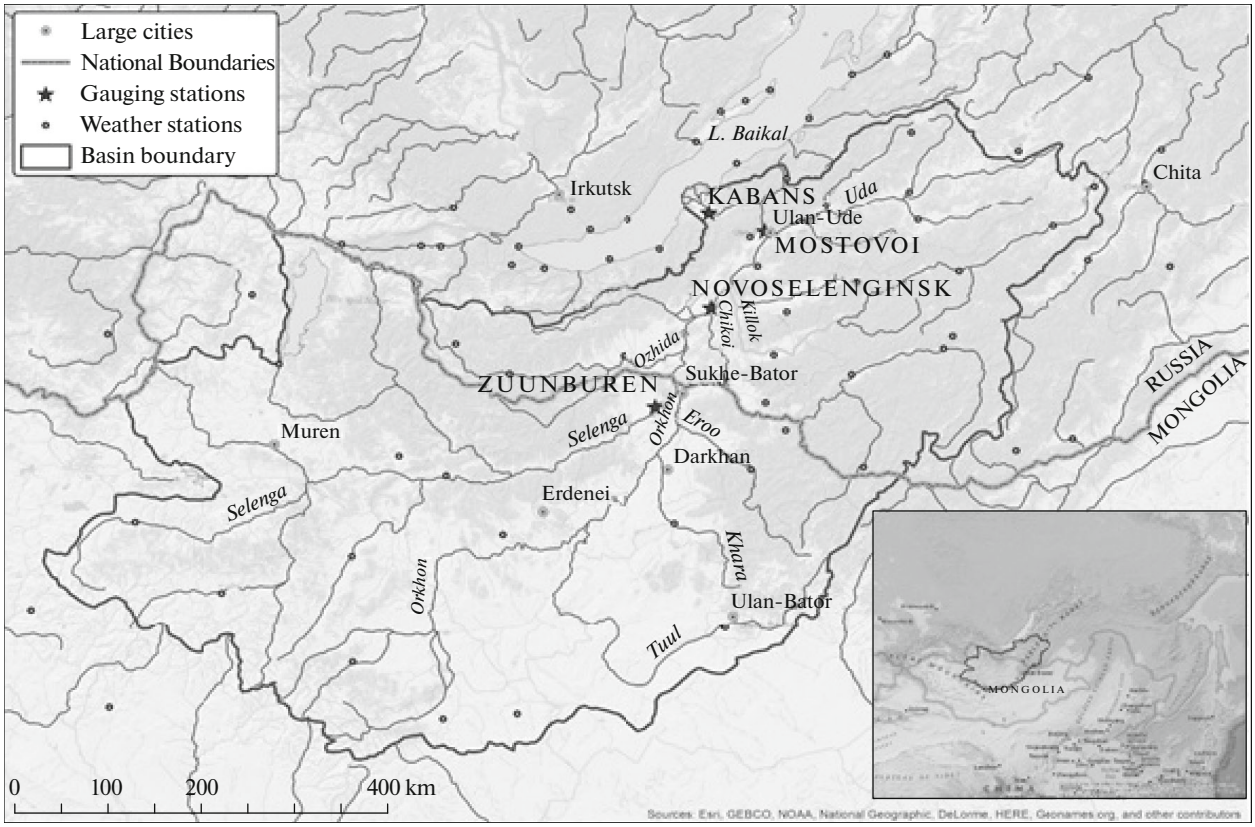


**Fig. 1.** Scheme of the Selenga River Basin with the locations of gauging stations, whose data were used in the study: Kabansk, Mostovoi, Novoselenginsk, Zuunburen.

surements of the basin's characteristics (topography, soil and vegetation properties).

In this study, the model parameters were set up using spatial data from global databases in web-GIS, in particular, digital elevation model USGS HydroSheds (equivalent to the spatial resolution of 1 × 1 km) [23], land-cover database GLCC (Global Land Cover Characterization) [25]. Due to the insufficiently detailed database for the Mongolian part of the basin, FAO Harmonized World Soil Database (HWSD) [9] was supplemented by specific Mongolian maps of soil types (Ecological Atlas of the Lake Baikal Basin [6]; the National Atlas of Mongolia [31]).

A list of ECOMAG parameters is given in [14]. From 3 to 5 key parameters are to be calibrated against streamflow data in different gauges and, if available, other hydrological variable observations (snow, soil moisture, groundwater level, etc.). The calibration procedure is described in [14].

The model is driven by time series of daily air temperature, air humidity and precipitation assigned from EWEMBI reanalysis dataset for the period of 1980–2013 on a 0.5° × 0.5° spatial grid. This reanalysis was carried out to support the bias correction in climate input data for the impact assessments carried out in Phase 2b of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP2b) [10].

The ECOMAG-based hydrological model of the Selenga River basin was first presented in [26]. In our study, the model was improved by using more reliable data on the underlying surface characteristics for the Mongolian part of the basin and updated meteorological reanalysis data. Additionally, model evaluation was carried out for longer streamflow observation series reflecting multi-year runoff decline tendency, which has been observed in the basin. The use of long-term observation series allows applying a sophisticated evaluation procedure [22] described in the next subsection.

### *Calibration/Evaluation Procedure*

The following procedure of the ECOMAG model calibration/evaluation, partly based on that suggested in [22], was carried out in this study:

(1) A modified version of the DSS-test [20] is used for calibration/evaluation to optimize the model simultaneously to periods with different climate.

(2) Model performance is evaluated for:

—multiple sites within the catchment to ensure internal consistency of the simulated processes;

—hydrological indicator of interest;

—observed trends.

The modified DSS-test [20] was designed as follows.

The model was calibrated against the Kabansk gauge streamflow observation data (daily, monthly, and annual) from January 1, 2000, to December 31, 2013. To represent the goodness of fit of the simulated and measured variables, we used an enhanced version (1), (3), (4) of the Nash-Sutcliffe (1970) efficiency criterion (1), (2).

$$\mathrm{NSE} = 1 - \frac{\sum_{j=1}^{N}\sum_{i=1}^{n}\left(Q^{j}_{f(i)} - Q^{j}_{s(i)}\right)^2}{\sum_{j=1}^{N}\sum_{i=1}^{n}\left(Q^{j}_{f(i)} - Q^{j}_{r(i)}\right)^2}, \quad (1)$$

where $Q^{j}_{f(i)}$ and $Q^{j}_{s(i)}$ are observed and simulated discharge values averaged over the $i$th time interval (day, month, year) of the $j$th year, respectively ($i = 1, 2, 3,\ldots, n$; $j = 1, 2, 3,\ldots, N$), $Q^{j}_{r(i)}$ is a discharge calculated from a reference (benchmark) model; $N$ is the number of observation years; $n$ is the number of time intervals within a year ($n = 365/366$ for the day, $n = 12$ for the month, $n = 1$ for the annual discharge).

As mentioned in the 2nd Section,

$$Q^{j}_{r(i)} = \overline{Q} = \frac{1}{N \times n}\sum_{j=1}^{N}\sum_{i=1}^{n} Q^{j}_{f(i)} \quad (2)$$

in the original Nash–Sutcliffe measure [30], and the hydrological model comparison with such primitive reference model (2) results in overestimation of the model efficiency. Taking this into account, we used more complex reference models describing either seasonal or inter-annual trends.

For daily and monthly discharge series, we used the seasonal trend-based reference model recommended in [12]:

$$Q^{j}_{r(i)} = Q_{\mathrm{season}(i)} = \frac{1}{N}\sum_{j=1}^{N} Q^{j}_{f(i)} \quad (3)$$

$i = 1, 2,\ldots, 365(366)$ or $12$; $j = 1, 2, 3,\ldots, N$.

For the annual discharge series, we modified the reference model as follows:

$$Q^{j}_{r(i)} = Q^{j}_{\mathrm{trend}} = a + bj, \quad (4)$$

where $a$ and $b$ are the coefficients of linear trend fitted to the annual discharge series; $i = 1$; $j = 1, 2, 3,\ldots, N$.

In order to distinguish the original NSE [30] from the trend-based one, we will use a denotation $\mathrm{NSE}_{s\_\mathrm{trend}}$ or $\mathrm{NSE}_{y\_\mathrm{trend}}$ for the efficiency measures used in the seasonal trend-based reference model (3) and the inter-annual trend-based reference model (4), respectively. If the trend-based efficiency measures are positive, then the model performance is acceptable.

The calibrated model was then evaluated for the periods of different climate observed in the basin. We divided the available meteorological data into four contrasting climate periods with regard to the ratios of the observed annual air temperature ($T^{j}_{\mathrm{annual}}$) and annual precipitation ($P^{j}_{\mathrm{annual}}$) to the corresponding mean annual values ($T_{cl}$ and $P_{cl}$, respectively): $T^{j}_{\mathrm{annual}} > T_{cl}$ and $P^{j}_{\mathrm{annual}} > P_{cl}$ (warm-wet, WW, period), $T^{j}_{\mathrm{annual}} < T_{cl}$ and $P^{j}_{\mathrm{annual}} > P_{cl}$ (cold-wet, CW, period), $T^{j}_{\mathrm{annual}} > T_{cl}$ and $P^{j}_{\mathrm{annual}} < P_{cl}$ (warm-dry, WD, period), and $T^{j}_{\mathrm{annual}} < T_{cl}$ and $P^{j}_{\mathrm{annual}} < P_{cl}$ (cold-dry, CD, period). (Note that all annual values were obtained from the used meteorological reanalysis data averaged over the catchment area). The WW period includes 6 years (1990, 1994, 1998, 2003, 2008, 2013); the CW period includes 10 years (1982, 1983, 1984, 1985, 1986, 1988, 1991, 1993, 2000, 2012); the WD period includes 11 years (1989, 1992, 1995, 1996, 1997, 1999, 2001, 2002, 2004, 2006, 2007), and the CD period includes 6 years (1980, 1981, 1987, 2005, 2010, 2011). The original NSE with the reference model (2) was used to assess the goodness of fit of the simulated and measured variables during the selected period. Thus, for each streamflow variable (daily, monthly, and annual), we calculated four NSE-estimations: $\mathrm{NSE}_1 = \mathrm{NSE}_{\mathrm{WW}}$ (NSE for the warm-wet period), $\mathrm{NSE}_2 = \mathrm{NSE}_{\mathrm{CW}}$ (NSE for the cold-wet period), $\mathrm{NSE}_3 = \mathrm{NSE}_{\mathrm{WD}}$ (NSE for the warm-dry period), and $\mathrm{NSE}_4 = \mathrm{NSE}_{\mathrm{CD}}$ (NSE for the cold-dry period). The model is considered as the robust one and can be chosen as a candidate for climate impact assessment if all differences $\mathrm{NSE}_i - \mathrm{NSE}_j$ ($i, j = 1, \ldots, 4$; $i > j$) are statistically insignificant. We proposed the following statistical test to compare the estimations.

The Nash–Sutcliffe efficiency measure NSE is a random variable depending on the statistics of the observed and simulated discharge series [29]. Let us assume that NSE is an normally distributed variable with a mean $M[\mathrm{NSE}]$ and a variance $\mathrm{Var}[\mathrm{NSE}]$. To test the null hypothesis that $\mathrm{NSE}_*$ in one population (one climate period) is the same as $\mathrm{NSE}_{**}$ in another population (another climate period), and assuming these variables as statistically independent, we introduce the following test statistic:

$$Z_{\mathrm{NSE}} = \frac{\left|M\left[\mathrm{NSE}_*\right] - M\left[\mathrm{NSE}_{**}\right]\right|}{\sqrt{\mathrm{Var}\left[\mathrm{NSE}_*\right] + \mathrm{Var}\left[\mathrm{NSE}_{**}\right]}}. \quad (5)$$

The null hypothesis is accepted with a confidence level of $(1 - \alpha)$ if

$$\Phi\left(Z_{\mathrm{NSE}}\right) < 1 - \alpha/2, \quad (6)$$

where $\Phi(x)$ is the cumulative distribution function of a standard normal distribution; $\alpha$ is a significance level (the probability of rejecting the null hypothesis when it is true).

Hereafter, the significance level is $\alpha = 0.05$.

Let us assume that for the $i$th climate period, $M[\mathrm{NSE}] = \mathrm{NSE}_i$ and $\mathrm{Var}[\mathrm{NSE}] = \sigma_i^2$. Then, according to (5) and (6), the difference between two independent estimations is statistically insignificant with a confidence level of $(1 - \alpha)$ if

$$\Phi(Z_{ij}) < 1 - \alpha/2 \quad (7)$$

where

$$Z_{ij} = \frac{|\mathrm{NSE}_i - \mathrm{NSE}_j|}{\sqrt{\sigma_i^2 + \sigma_j^2}} \quad (8)$$

$i,j = 1, ...,4$; $i > j$.

Sample NSE-variance for $i$th climate period is defined as

$$\sigma_{\mathrm{NSE}}^2 = \frac{4}{N}\left[(\rho - \alpha)^2\alpha^2 + N\alpha^2\left(\frac{1-\rho^2}{N-2}\right) + \beta^2\left(\alpha^2 + \beta^2 + 1\right)\right], \quad (9)$$

where $\rho$ is Pearson's correlation coefficient between the simulated and observed discharge series; $\alpha = \frac{\sigma_s}{\sigma_f}$, $\sigma_s$, and $\sigma_f$ are the standard deviations of the simulated and observed discharge series, respectively; $\beta = \frac{\mu_s - \mu_f}{\sigma_f}$; $\mu_s$ and $\mu_f$ are the mean values for the simulated and observed discharge series, respectively; $N$ is sample size.

The proof of the expression (9) based on the NSE decomposition [16, 29] and first-order Taylor series linearization is beyond the framework of the paper.

Thus, the modified DSS-test is assumed to be successfully passed for a given hydrological indicator (daily, monthly, or annual discharge) if and only if the condition (7) is satisfied for all differences between the NSE-values estimated for four contrast periods, i.e., for all six combinations of $\mathrm{NSE}_i - \mathrm{NSE}_j$ ($i, j = 1, ... , 4$; $i > j$).

We applied the modified DSS-test to evaluate the ECOMAG model performance for four different sites within the catchment (Kabansk, Novoselenginsk, Mostovoi, Zuunburen; see Fig. 1) and to three hydrological indicators (daily, monthly, and annual discharge).

Finally, trends in the observed and simulated streamflow series were compared and analyzed for the specified gauging stations and hydrological indicators.

## RESULTS AND DISCUSSION

### *Calibration Results*

Figure 2 shows observed historic and simulated hydrographs (daily, monthly, and annual discharge) at the outlet gauging station Kabansk over the calibration period (2000–2013). Trend-based efficiency criteria estimated for four gauge stations located within the Selenga River catchment (the whole simulation period 1980–2013) are given in Table 1. For comparison, the standard Nash–Sutcliffe efficiency criteria are also given in the table. As one can see from Table 1, the model performance is acceptable in terms of the enhanced trend-based efficiency measures: both $\mathrm{NSE}_{s\_\mathrm{trend}}$ and $\mathrm{NSE}_{y\_\mathrm{trend}}$ appeared to be positive. This means that, with the optimal values of the model parameters adjusted against the observation data at the outlet gauging station for the period 2000–2013, the model correctly reproduced long-term historical runoff data (daily, monthly, and annual discharge) at four sites within the Selenga River Basin. The model efficiency turned out to be approximately the same for the three gauging stations located in the lower reaches of the river, but it was significantly lower for the gauging station Zuunburen located in the middle reaches of the river; in other words, the model reproduced the water regime in the upstream (Mongolian) part with a lesser accuracy. This result can be explained by the lower reliability of reanalysis data in this part of the basin due to a much sparser meteorological station network and a less detailed database on the underlying surface than in the Russian part of the basin.

The obtained results confirmed that the use of the trend-based efficiency measure sharply raises the requirements for the hydrological model under test than the standard NSE criterion. This finding is more noticeable for daily discharges: the average trend-based efficiency is 0.37 lower than the standard NSE. For average monthly discharges, the average efficiency is 0.36 lower than NSE. The decrease in efficiency is due to the fact that the trend-based reference models describe a significant portion of the observed discharge variability.

There is also a small decrease (0.04) in the $\mathrm{NSE}_{y\_\mathrm{trend}}$ values compared to the corresponding NSE values for annual discharge series. This means that in 1980–2013, a linear trend describing part of the variance appears in these series (Fig. 3).

### *Model Evaluation for Contrasting Climate Periods*

NSE measures estimated by formulas (1), (2) for the chosen contrasting climate periods, as well as the sample standard deviations (Eq. (9)) of these measures are shown in Table 2. Additionally, calculations of the test statistics $Z_{ij}$ are presented in this table. Overall, in Table 2 gives data for verification of the inequality (7), which is the necessary condition for the successful completion of the modified DSS-test.

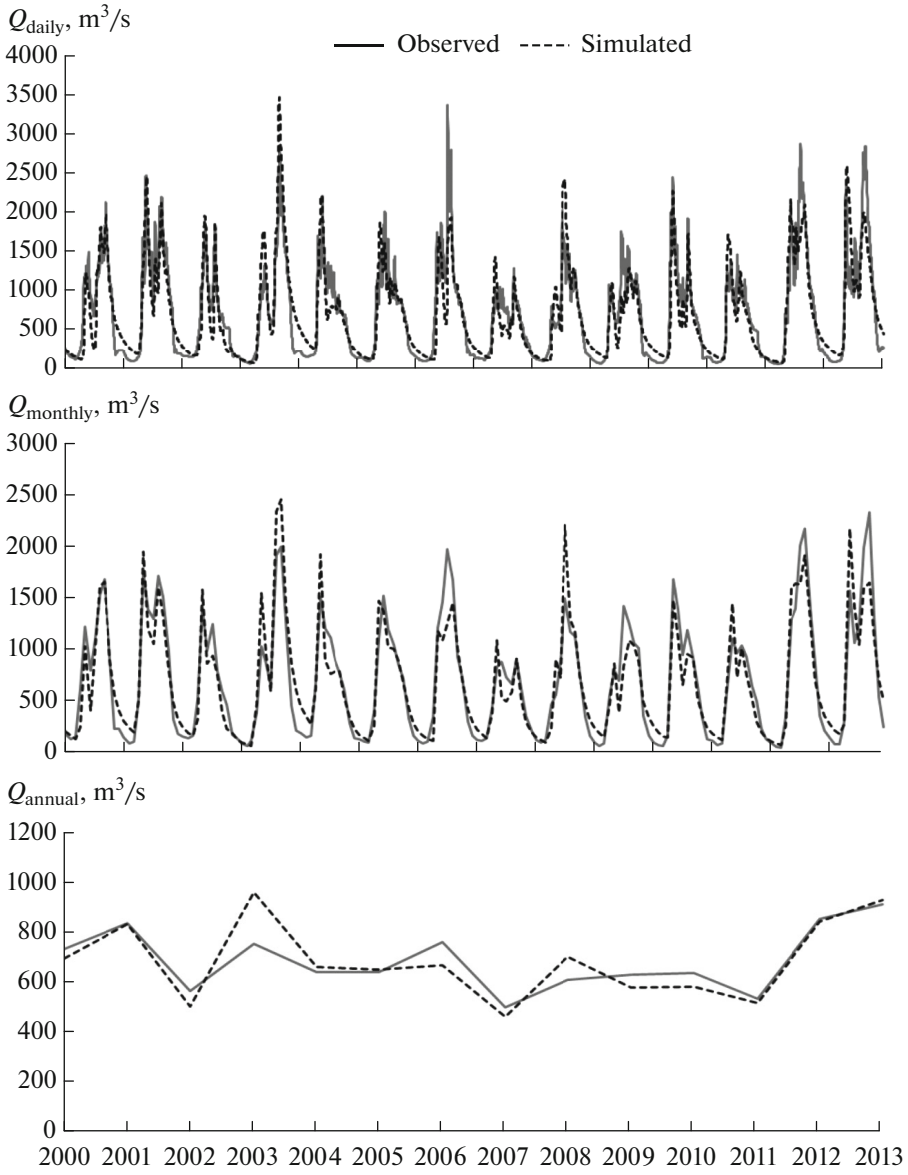


**Fig. 2.** Observed and simulated hydrographs (daily, monthly, and annual discharge) for the calibration period 2000–2013 at the outlet Kabansk gauging station.

The above results show that the model successfully passed the developed test for average monthly and annual discharges at all four gauging stations. In other words, the model demonstrated statistically similar performance for contrasting climate periods, a fact which indicates its sufficient robustness for climate impact assessment. At the same time, though the difference in the NSE-values estimated for daily discharge is not too large (for instance, the maximum difference is only 0.11 for the basin outlet Kabansk; see Table 2), but the DSS-test had not been passed, i.e., the model appeared to be insufficiently robust for the daily data. The reason is the small NSE sample variance, which, in turn, is due to the large sample size for daily discharge series.

### *Trend Analysis*

The slope of linear trend lines fitted to the observed and simulated discharge series of the Selenga River at four gauging stations are compared in Table 3. Figure 3 exemplifies the annual runoff series. We found the same tendency to exist in the simulated and observed series, namely, a decrease in runoff over the study period at four gauging stations. According to both observations and simulations, the decrease in the summer and autumn runoff is most rapid, whereas

**Table 1.** Trend-based efficiency measures ($NSE_{s\_trend}$ and $NSE_{y\_trend}$) estimated for four gauge stations located within the Selenga River catchment (the whole simulation period 1980–2013). The efficiency measure NSE (Eqs. (1), (2)) is shown in brackets

| Gauging station | Catchment area, km$^2$ | $NSE_{s\_trend}$ (Eqs. (1), (3)) | | $NSE_{y\_trend}$ (Eqs. (1), (4)) |
|---|---|---|---|---|
| | | daily discharge | monthly discharge | annual discharge |
| Kabansk | 445000 | 0.46 (0.81) | 0.51 (0.84) | 0.74 (0.79) |
| Novoselenginsk | 440000 | 0.47 (0.80) | 0.52 (0.85) | 0.71 (0.77) |
| Mostovoi | 360000 | 0.38 (0.79) | 0.50 (0.84) | 0.76 (0.80) |
| Zuunburen | 148000 | 0.22 (0.59) | 0.22 (0.65) | 0.65 (0.68) |

**Table 2.** Nash–Sutcliffe Efficiency (NSE) values (formulas (1)–(2)), their sample standard deviations ($\sigma_{NSE_*}$) (formula (9)) and test statistics (formula (5)) estimated for the contrasting climate periods. Shared cells denote test-statistics for which the conditions (7) has not been met

| Climate period | Statistics | Kabansk | | | Novoselenginsk | | | Mostovoi | | | Zuunburen | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | daily | monthly | annual | daily | monthly | annual | daily | monthly | annual | daily | monthly | annual |
| | | Nash–Sutcliffe Efficiency NSE | | | | | | | | | | | |
| WW | $NSE_1$ | 0.86 | 0.89 | 0.80 | 0.85 | 0.88 | 0.75 | 0.83 | 0.86 | 0.67 | 0.62 | 0.63 | –2.52 |
| CW | $NSE_2$ | 0.79 | 0.83 | 0.61 | 0.81 | 0.86 | 0.71 | 0.80 | 0.85 | 0.77 | 0.66 | 0.70 | 0.70 |
| WD | $NSE_3$ | 0.75 | 0.78 | 0.55 | 0.74 | 0.78 | 0.64 | 0.73 | 0.78 | 0.66 | 0.44 | 0.48 | 0.38 |
| CD | $NSE_4$ | 0.78 | 0.84 | –0.40 | 0.73 | 0.79 | –2.64 | 0.74 | 0.81 | –1.76 | 0.19 | 0.19 | –3.59 |
| | | Sample standard deviation of NSE (formula (9)) | | | | | | | | | | | |
| WW | $\sigma_1$ | 0.015 | 0.079 | 0.466 | 0.016 | 0.081 | 0.533 | 0.017 | 0.090 | 0.605 | 0.032 | 0.187 | 4.118 |
| CW | $\sigma_2$ | 0.013 | 0.065 | 0.486 | 0.012 | 0.060 | 0.435 | 0.012 | 0.060 | 0.380 | 0.017 | 0.093 | 0.437 |
| WD | $\sigma_3$ | 0.012 | 0.065 | 0.497 | 0.012 | 0.063 | 0.426 | 0.012 | 0.064 | 0.397 | 0.023 | 0.128 | 0.681 |
| CD | $\sigma_4$ | 0.020 | 0.093 | 1.953 | 0.022 | 0.109 | 4.324 | 0.021 | 0.099 | 3.078 | 0.168 | 0.283 | 4.526 |
| | | Test statistic (formula (5)) | | | | | | | | | | | |
| | $Z_{21}$ | 3.527 | 0.586 | 0.282 | 2.000 | 0.198 | 0.058 | 1.442 | 0.092 | 0.140 | 1.104 | 0.335 | 0.778 |
| | $Z_{31}$ | 5.726 | 1.075 | 0.367 | 5.500 | 0.975 | 0.161 | 4.806 | 0.724 | 0.014 | 4.568 | 0.662 | 0.695 |
| | $Z_{41}$ | 3.200 | 0.410 | 0.598 | 4.411 | 0.663 | 0.778 | 3.331 | 0.374 | 0.775 | 2.514 | 1.297 | 0.175 |
| | $Z_{32}$ | 2.261 | 0.544 | 0.086 | 4.125 | 0.920 | 0.115 | 4.125 | 0.798 | 0.200 | 7.692 | 1.390 | 0.395 |
| | $Z_{42}$ | 0.419 | 0.088 | 0.502 | 3.192 | 0.563 | 0.771 | 2.481 | 0.346 | 0.816 | 2.783 | 1.712 | 0.943 |
| | $Z_{43}$ | 1.286 | 0.529 | 0.471 | 0.399 | 0.079 | 0.755 | 0.413 | 0.254 | 0.780 | 1.474 | 0.934 | 0.867 |
| | | Has the DSS test been passed? | | | | | | | | | | | |
| | $\Phi(Z_{ij}) < 1 - \alpha/2$ $(i,j = 1, ..., 4; i > j)$ | No | Yes | Yes | No | Yes | Yes | No | Yes | Yes | No | Yes | Yes |

that in winter and spring runoff, conversely, is slow. Both observed and simulated runoff at the downstream gauging stations in the Selenga River decrease more rapidly than at the middle-reach gauging stations. The summer runoff decrease (the largest among all seasons) is reproduced well by the model: for the observed and simulated runoff series, it is about 200–210 m$^3$/s per decade. Note that summer runoff makes the largest contribution to the annual runoff of the Selenga River [11]. One can see from Table 3 that, in

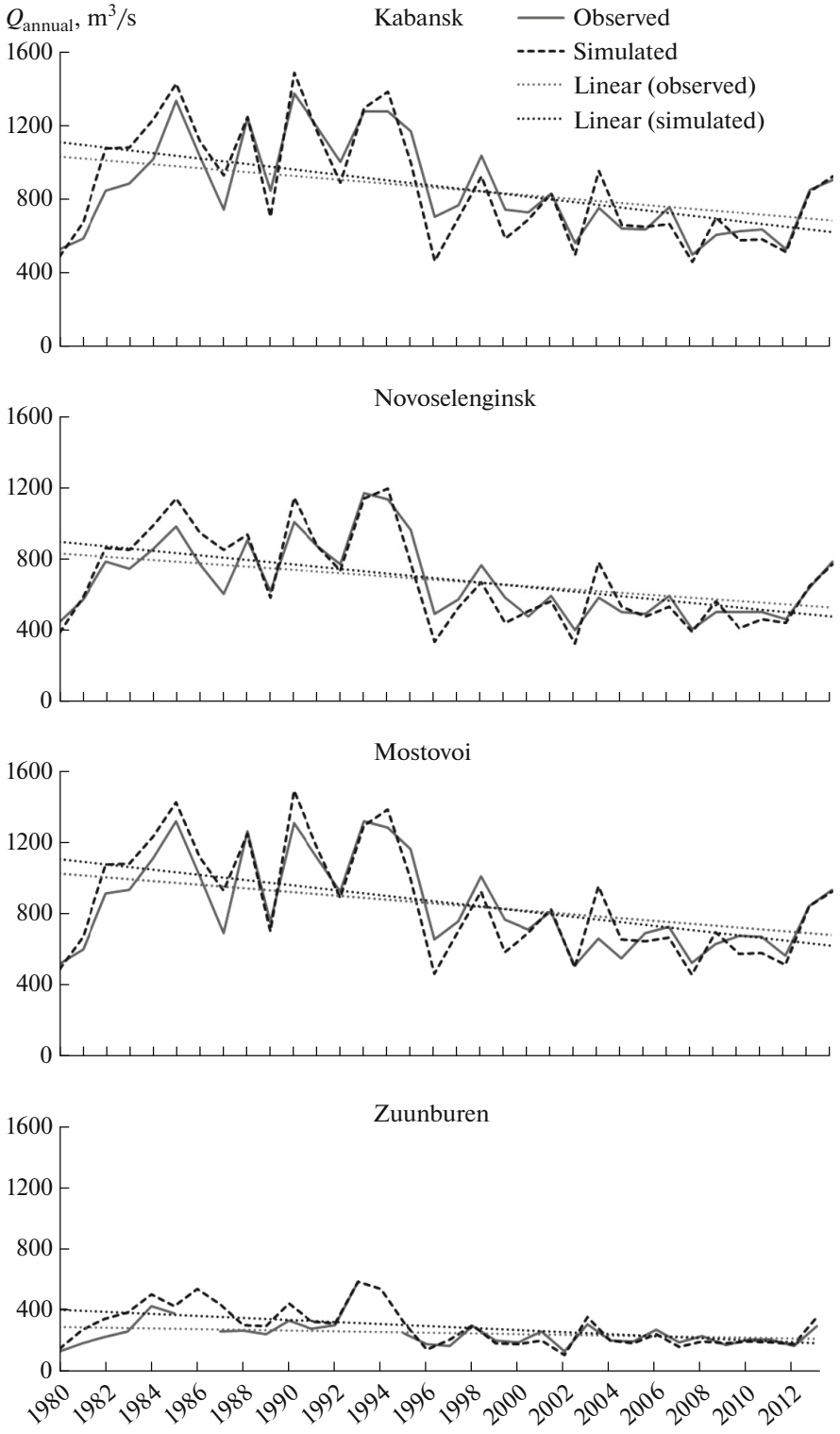


**Fig. 3.** Observed and simulated annual discharge series (with linear trends) for the period of 1980–2013 at four sites within the Selenga River Basin.

most cases, the decline rates are overestimated by the model. The largest errors were obtained for the winter and spring seasons, where they were, to some extent, due to the very small absolute values of the trend slope for these seasons. According to the annual runoff observations, decreasing trend slope is approximately

**Table 3.** Slope of trend lines ($m^3$/s per year) fitted to the observed and simulated discharge series

| Gauging station | Annual discharge | | Mean spring discharge | | Mean summer discharge | | Mean autumn discharge | | Mean winter discharge | |
|---|---|---|---|---|---|---|---|---|---|---|
| | observed | simulated | observed | simulated | observed | simulated | observed | simulated | observed | simulated |
| Kabansk | −10.60 | −14.71 | −7.11 | −1.31 | −28.29 | −24.53 | −22.28 | −35.65 | −0.53 | −9.47 |
| Novoselenginsk | −9.06 | −12.54 | −7.01 | −1.92 | −21.86 | −24.61 | −17.85 | −28.71 | −0.28 | −8.09 |
| Mostovoi | −10.42 | −14.69 | −1.28 | −5.11 | −28.98 | −24.59 | −22.45 | −35.60 | −1.12 | −9.45 |
| Zuunburen | −2.35 | −5.77 | −1.22 | −0.25 | −5.81 | −7.74 | −5.49 | −12.97 | −0.89 | −5.04 |
| Mean slope | −8.11 | −11.93 | −4.16 | −2.15 | −21.24 | −20.37 | −17.02 | −28.23 | −0.71 | −8.01 |

80 $m^3$/s per decade whereas it is approximately 120 $m^3$/s per decade for the corresponding simulation series.

## CONCLUSIONS

*All models look good, but some are useless*
*<for impact studies>*
*(a rephrase of George Box's quote).*

In the recent decades, there is a large gap between the progress in improving hydrological models, the methods of their calibration, their ability to assimilate different types of data, etc., on the one hand, and the outdated methods of model evaluation and testing, on the other hand. We still use easy-to-pass tests and very "soft" statistical criteria of model efficiency. As a result, there are a lot of "good" models that pretend to be suitable for impact studies, and the number of such models grows like a snowball. However, though the performance of all models is mostly good, but, most likely, some of them are useless. In order not to be overwhelmed by pseudo-good models and to understand the grounds for credibility of a given hydrological model, we have to be able to distinguish between models appropriate for impact studies and unsuitable ones.

It seems reasonable to search these grounds within the framework of the following pragmatic argumentation, which is based on three well-established statements [1, 5, 20, 33, 38]. First, a predictive hydrological model can never be universally validated, but its performance can be evaluated for situations that imitate the "target" conditions of the model application. Second, if a model does not perform well, this implies that the model is most likely inadequate under the "target" conditions. Third, the opposite is not true: the lack of disagreement does not necessarily imply the model applicability for these conditions; however, appropriate evaluation design increases the credibility of and decreases the uncertainty in the model results.

As the model's predictive ability usually cannot be tested directly with the use of data (the latter, most probably, never be available), specific test ("crash-test") is necessary, allowing one to reinforce the model's applicability.

In this study, we suggested such a crash test, founded partly on the calibration/validation procedure presented in [22], and applied the test to evaluating the performance of the physically based model of the Selenga River runoff generation. First, for calibration, we established an enhanced NSE criterion including models (3) and (4), which are more complex and trend-oriented reference (benchmark) than the simple reference model (2) used in the original NSE criterion. Second, we introduced a modified DSS-test focused on comparing the model performance criteria for climatically contrasting periods, extracted from the observation data, and proposed a new statistical measure (7)–(8) for obtaining the DSS-test result. Third, the model's performance was evaluated for four sites within the catchment using three indicators of interest (daily, monthly, and annual discharge series). The model's ability to reproduce the observed trends in annual and seasonal discharge values was assessed.

Our findings can be summarized as follows:

(1) The model is demonstrated to be effective in terms of the trend-oriented NSE measures, which impose higher requirements on the tested model than the ordinary NSE. This result is obtained for all considered hydrological variables and gauging stations.

(2) The differences in the NSE measures obtained for the contrasting climate periods appeared statistically insignificant in terms of monthly and annual discharge series for all considered gauging stations, i.e. the model successfully passed the modified DSS-test for these hydrological indicators. However, the model has not passed the DSS-test in terms of the daily discharge series because of small sample variance of the NSE measure for this series. Thus, we conclude that the model is robust enough to be applied to assess climate change impact on the annual and monthly runoff in different parts of the Selenga River basin.

(3) The model demonstrates its ability to reproduce the decreasing trend in multiyear variations of daily, monthly, and annual discharge series observed at all considered gauging stations. In accordance with observations, the interannual decrease in the simu-

lated summer and autumn runoff is more intensive than the trends in other seasons. Also, the Lower Selenga runoff decreases faster than that in the Middle Selenga for both observations and simulations. The decreasing trend in the summer runoff (about 200–210 $m^3/s$ per decade) is well simulated by the model. The errors in the trend slope are highest for winter and spring runoff, but the absolute trend values are negligible for these seasons. Overall, the model overestimates trend slope comparing with observations.

## ACKNOWLEDGEMENTS

The development of Selenga hydrological model was supported by the Russian Science Foundation, project no. 14-17-00700P. The development of the calibration/evaluation procedure was supported by the Russian Science Foundation, project no. 17-77-3006. The creation of the databases and GIS technologies for modeling was supported by the Russian Foundation for Basic Research, project no. 17-29-05027.

The present work was carried out under the Panta Rhei Research Initiative of the International Association of Hydrological Sciences (IAHS).

## REFERENCES

1. Andréassian, V., Le Moine N., Perrin C., Ramos M.-H., Oudin L., Mathevet T., Lerat J., and Berthet L., All that glitters is not gold: The case of calibrating hydrological models, *Hydrol. Processes*, 2012, vol. 26, no. 14, pp. 2206–2210. doi 10.1002/hyp.9264
2. Beven, K., Towards a coherent philosophy for modelling the environment, *Proc. R. Soc. London*, 2002, Ser. A, vol. 458, pp. 2465–2484.
3. Chalov, S.R., Jarsjö, J., Kasimov, N.S., Romanchenko, A., Pietron´, J., Thorslund, J., and Belazerova, E., Spatiotemporal variation of suspended transport in the Selenga Basin (Mongolia and Russia), *Environ. Earth Sci.*, 2014, vol. 73, no. 2, pp. 663–680. doi 10.1007/s12665-014-3106-z
4. Chalov, S.R., Thorslund, J., Kasimov, N., Nittrouer, J., Iliyecheva, E., Pietron, J., Shinkareva, G., Lychagin, M., Aybullatov, D., Kositky. A., Tarasov, M., Akhtman, Y., Garmaev, E., Karthe, D., and Jarsjo, J., The Selenga River Delta—Geochemical barrier for protecting Lake Baikal's waters, *Regional Environ. Change*, 2016, vol. 17, no. 7, pp. 2039–2053. doi 10.1007/s10113-016-0996-1
5. Coron, L., Andréassian, V., Bourqui, M., Perrin, C., and Hendrickx, F., Pathologies of hydrological models used in changing climatic conditions: a review, *IAHS Publ.*, 2011, vol., 344, pp. 39–44.
6. *Ecological atlas of the Lake Baikal basin*: Irkutsk: Institute of Geography, Sib. Branch, Russ. Acad. Sci., 2014.
7. Euser, T, Winsemius, H.C., Hrachowitz, M., Fenicia, F., Uhlenbrook, S., and Savenije, H.G., A framework to assess the realism of model structures using hydrological signatures, *Hydrol. Earth Syst. Sci.*, 2013, vol. 17, pp. 1893–1912. doi 10.5194/hess-17-1893-2013
8. Ewen, J. and Parkin, G., Validation of catchment models for predicting land-use and climate change impacts. 1, Method. *J. Hydrol.*, 1996, vol. 175, pp. 583–594.
9. *FAO/IIASA/ISRIC/ISS-CAS/JRC, Harmonized World Soil Database (version 1.2)*, Rome-Laxenburg: FAO, 2012.
10. Frieler, K., Lange, S., Piontek, F., Reyer, C.P.O., Schewe, J., Warszawski, L., Zhao, F., Chini, L., Denvil, S., Emanuel, K., Geiger, T., Halladay, K., Hurtt, G., Mengel, M., Murakami, D., Ostberg, S., Popp, A., Riva, R., Stevanovic, M., Suzuki, T., Volkholz, J., Burke, E., Ciais, P., Ebi, K., Eddy, T.D., Elliott, J., Galbraith, E., Gosling, S.N, Hattermann, F., Hickler, T., Hinkel, J., Hof, C., Huber, V., Jägermeyr, J., Krysanova, V., Marcé, R., Müller Schmied, M., Mouratiadou, I., Pierson, D., Tittensor, D.P., Vautard, R., van Vliet, M., Biber, M.F., Betts, R.A., Bodirsky, B.L., Deryng, D., Frolking, S., D. Jones, C.D., Lotze, H.K., Lotze-Campen, H., Sahajpal, R., Thonicke, K., Tian, H., and Yamagata, Y., Assessing the impacts of 1.5°C global warming—simulation protocol of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP2b), *Geosci. Model Dev.*, 2017, vol. 10, no. 12, pp. 4321–4345. doi 10.5194/gmd-10-4321-2017
11. Frolova, N.L., Belyakova, P.A., Grigor'ev, V.Yu., Sazonov, A.A., and Zotov, L.V., Many-Year Variations of River Runoff in the Selenga Basin, *Water Resour.*, 2017, vol. 44, no. 3, pp. 243–255.
12. Garrick M., Cunnane C., and Nash J.E., A criterion of efficiency for rainfall-runoff models, *J. Hydrol.*, 1978, vol. 36, no. 3–4, pp. 375–381.
13. Gelfan, A., Gustafsson, D., Motovilov, Y., Arheimer, B., Kalugin, A., Krylenko, I., and Lavrenov, A., Climate change impact on water regime of two great arctic rivers: modeling and uncertainty issues, *Clim. Change*, 2017, vol. 141, no. 3, pp. 499–515. doi 10.1007/s10584-016-1710-5
14. Gelfan, A., Motovilov, Yu., Krylenko, I., Moreido, V., and Zakharova, E., Testing the robustness of the physically-based ECOMAG model with respect to changing conditions, *Hydrol. Sci. J.*, 2015, vol. 60, pp. 1266–1285. doi 10.1080/02626667.2014.935780
15. Gottschalk. L., Beldring. S., Engeland. K., Tallaksen. L., Sælthun. N.R., Kolberg. S., and Motovilov. Yu., Regional/macroscale hydrological modelling: A Scandinavian experience, *Hydrol. Sci. J.*, 2001, vol. 46, pp. 963–982.
16. Gupta H.V., Kling H., Yilmaz K.K., and Martinez G.F., Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, *J. Hydrol.*, 2009, vol. 377, pp. 80–91. doi 10.1016/j.jhydrol.2009.08.003
17. Huang, S., Kumar, R., Flörke, M., Yang, T., Hundecha, Y., Kraft, P., Gao, C., Gelfan, A., Liersch, S., Lobanova, A., Strauch, M., van Ogtrop, F., Reinhardt, J., Haberlandt, U., and Krysanova, V., Evaluation of an ensemble of regional hydrological models in 12 large-scale river basins worldwide, *Clim. Change*, 2017, vol. 141, pp. 381–397. doi 10.1007/s10584-016-1841-8

18. Kalugin, A.S. and Motovilov, Yu.G., Runoff formation model for the Amur river basin, *Water Resour.*, 2018, vol. 45, no. 2, pp. 149–159.

19. Kirchner, J.W., Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology, *Water Resour. Res.*, 2006, vol. 42, W03S04. doi 10.1029/2005WR004362

20. Klemeš, V. Operational testing of hydrological simulation models, *Hydrol. Sci. J.*, 1986, vol. 31, pp. 13–24.

21. Konikow, L.F. and Bredehoeft, J.D., Groundwater models cannot be validated, *Adv. Water Resour.*, 1992, vol, 15, pp. 47–62.

22. Krysanova, V., Donnelly, C., Gelfan, A., Gerten, D., Arheimer, B., Hattermann, F., and Kundzewicz, Z.W., How the performance of hydrological models relates to credibility of projections under climate change, *Hydrol. Sci. J.*, 2018, vol. 63, pp. 696–720. doi 10.1080/02626667.2018.1446214

23. Lehner, B., Verdin, K., and Jarvis., A., New global hydrography derived from spaceborne elevation data, *Eos Trans.*, 2008, vol. 89, no. 10, pp. 93–94.

24. Logachev, N.A., History and geodynamics of the Baikal rift, *Russ. Geol. Geophys.*, 2003, vol. 44, no. 5, pp. 391–406.

25. Loveland, T.R., Reed, B.C., Brown, J.F., Ohlen, D.O., Zhu, Z., Yang, L., and Merchant, J.W., Development of a global landcover characteristics database and IGBP DISCover from 1 km AVHRR data, *Int. J. Remote Sens.*, 2000, vol. 21, pp. 1303–1330.

26. Moreido, V.M. and Kalugin, A.S., Assessing possible changes in Selenga R. water regime in the XXI century based on a runoff formation model, *Water Resour.*, 2017, vol. 44, no. 3, pp. 390–398.

27. Motovilov, Yu.G., Hydrological simulation of river basins at different spatial scales: 1. generalization and averaging algorithms, *Water Resour.*, 2016, vol. 43, no. 3, pp. 429–437.

28. Motovilov, Yu., Gottschalk, L., Engeland, L., and Rodhe A., Validation of a distributed hydrological model against spatial observation, *Agric. Forest Meteor.*, 1999, vols. 98–99, pp. 257–277.

29. Murphy A.H., Skill scores based on the mean square error and their relationships to the correlation coefficient, *Mon. Weather Rev.*, 1988, vol. 116, pp. 2417–2424.

30. Nash, J.E. and Sutcliffe, J.V., River flow forecasting through conceptual models, Part I—A discussion of principles, *J. Hydrol.*, 1970, vol. 10, pp. 282–290.

31. *The National Atlas of Mongolia*, The Institute of Geography of Mongolian Academy of Science, 2009.

32. Oreskes, N., The role of quantitative models in science, in *Models in ecosystem science*, Princeton University Press, 2003, pp. 13–31.

33. Refsgaard, J.C., Madsen, H., Andréassian, V., Arnbjerg-Nielsen, K., Davidson, T.A., Drews, M., Hamilton, D.P., Jeppesen, E., Kjellström, E., Olesen, J.E., Sonnenborg, T.O., Trolle, D., Willems, P., and Christensen, J.H., A framework for testing the ability of models to project climate change and its impacts, *Clim. Change*, 2013, vol. 122, pp. 271–282.

34. Schaefli, B, Hingray, B, Niggli, M, and Musy, A., A conceptual glaciohydrological model for high mountainous catchments, *Hydrol. Earth Syst. Sci.*, 2005, vol. 9, pp. 95–109.

35. Seibert, J., On the need for benchmarks in hydrological modelling, *Hydrol. Processes*, 2001, vol. 15, pp. 1063–1064. doi 10.1002/hyp.446

36. Seibert, J., Reliability of model predictions outside calibration conditions, *Nord. Hydrol.*, 2003, vol. 34, pp. 477–492.

37. Thirel, G., Andréassian V., and Perrin C., On the need to test hydrological models under changing conditions, *Hydrol. Sci. J.*, 2015, vol. 60, nos. 7–8, pp. 1165-1173. doi 10.1080/02626667.2015.1050027

38. Thirel, G., Andréassian, V., Perrin, C., Audouy, J.-N., Berthet, L., Folton, N., Furusho, C., Kuentz, A., Lerat, J., Lindström, G., Martin, E., Mathevet, T., Merz, R., Parajka, J., Ruelland, D., and Vaze, J., Hydrology under change: an evaluation protocol to investigate how hydrological models deal with changing catchments, *Hydrol. Sci. J.*, 2015, vol. 60, nos. 7–8, pp. 1184–1199. doi 10.1080/02626667.2014.967248

39. Tornqvist, R., Jarsjo, J., Pietron, J., Bring, A., Rogberg, P., Asokan, S.M., and Destouni, G., Evolution of the hydro-climate system in the Lake Baikal basin, *J. Hydrol.*, 2015, vol. 519, pp. 1953–1962.

40. *World Meteorological Organisation. Intercomparison of Models of Snowmelt Runoff*, Operational Hydrology Report No. 23. Secretariat of the World Meteorological Organization, Geneva, Switzerland, 1986.

41. Wöhling, T., Samaniego, L., and Kumar, R., Evaluating multiple performance criteria to calibrate the distributed hydrological model of the upper Neckar catchment, *Environ. Earth Sci.*, 2013, vol. 69, pp. 453–468. doi 10.1007/s12665-013-2306-2

42. Yilmaz, K.K., Gupta, H.V., and Wagener, T., A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, *Water Resour. Res.*, 2008, vol. 44, W09417. doi 10.1029/2007WR006716