New Approaches to Hydrological Prediction in Data-sparse Regions (Proc. of Symposium HS.2 at the Joint IAHS & IAH Convention, Hyderabad, India, September 2009). IAHS Publ. 333, 2009.

Using recently developed global data sets for hydrological predictions

JOHAN STRÖMQVIST, JOEL DAHNE, CHANTAL DONNELLY, GÖRAN LINDSTRÖM, JÖRGEN ROSBERG, CHARLOTTA PERS, WEI YANG & BERIT ARHEIMER

Swedish Meteorological and Hydrological Institute (SMHI), SE-601 76 Norrköping, Sweden
johan.stromqvist@smhi.se

Abstract The HYPE hydrological model was used for multi-basin applications with model input derived from global databases compiled using the World Hydrological Input Set-up Tool (WHIST). The model was applied to the La Plata Basin (3.2 million km^2) in South America and to Europe (7 million km^2). Water balance was modelled reasonably well, with volume errors at the gauging stations in Europe being generally <10%, whilst there were larger discrepancies in La Plata Basin. The median Nash-Sutcliffe model efficiency (NSE) was 0.27 for Europe and <0 for La Plata Basin. A simple sensitivity study shows that, for Northern Europe, the model results were most sensitive to meteorological forcing data and land cover. The results indicate that global databases can be useful for hydrological predictions in data sparse regions, although further studies are required to better distinguish between specific sources of errors and possibilities for improvements of both databases and models.

Key words hydrological modelling; predictions in ungauged basins; sensitivity analysis; global databases

INTRODUCTION

Availability of input data for hydrological predictions can be limited due to lack of accessibility, low reliability or insufficient monitoring efforts. Important basic data needs in hydrological modelling normally include long-term meteorological data, soil and land cover data, and information on river basin hydrography. In addition, hydrological observations are needed for calibration and validation. The magnitude of the issues with data availability that need to be addressed varies between countries and regions, and depends both on the type of data required by a specific model, and on the purpose of the hydrological prediction. Lack of input data is not limited to the developing world, and is particularly notable for transboundary river basins, which are very common in, for example, Europe. At present, large-scale hydrological predictions are required for more homogenous forecasting and early warning systems, integrated water management and adaptation strategies to climate change.

An approach for large-scale hydrological predictions in which global data sets are used as a substitute for, and complement to, local data is therefore presented here. The approach can be defined as a system of methods for automatic generation of model input data and is called WHIST (World Hydrological Input Set-up Tool). It consists of a number of FORTRAN and JAVA programs with the specific function to organize the input data to be applicable for multi-basin hydrological modelling. The input data include both basin physiography and model forcing and it has so far been used for applications in the Arctic, Europe and South America. In this paper, we have used the WHIST for generation of input data files to the Hydrological Predictions for the Environment (HYPE) model (Arheimer *et al*., 2008; Lindström *et al*., 2009). The aim is to: (1) quantify the value of using global data sets for predictions in ungauged basins; and (2) identify critical input data sets and their limitations which may affect the results of the hydrological modelling. The results of large-scale multi-basin modelling are compared to monitored time series at independent sites, and a sensitivity analysis is performed in one basin.

MATERIAL AND METHODS

WHIST and global databases

An increasing number of global data sets with seamless data are becoming readily available. Data sets that so far have been used within the WHIST framework are listed in Table 1. The system

Copyright © 2009 IAHS Press

Table 1 Global data sets used for large-scale hydrological predictions in data sparse regions.

Global data set	Database	Resolution	Source
Topographic data	Hydro1k	1 km	USGeological Survey (USGS)
Land cover	ECOCLIMAP	Approx. 1 km	Champeaux (2005)
Soil type	Soil map of the world	10 km	UN Educational,Scientific and Cultural Organization (UNESCO)
Precipitation and temperature:			
1957–2002	ERA-40	1º	European Centre for Medium-Range Weather Forecasts (ECMWF)
1989–2007	ERA-interim	0.75º	ECMWF
Water discharge	3700 gauging stations	Daily/monthly	Global Runoff Data Centre (GRDC), Koblenz (GRDC, 2008)

makes use of a hydrologically corrected gridded topographic database, Hydro1k (USGS, 2000), for automatic delineation of river basins. The delineation is either made for user specified geographical locations (e.g. river gauging stations and river branches) or for whole coastal stretches. In the routine, a list with all Hydro1k grid cells upstream of the specified flow points is created. These cells are found by using the information on grid cell flow direction in Hydro1k. Land cover and soil data information for each of the included grid cells are abstracted from the ECOCLIMAP and the UNESCO soil map of the world, respectively. Similar land covers are grouped into classes and soils are grouped into texture based soil classes.

River basins are further divided into sub-basin of relevant size for the application (determined by the user) and the basin routing order is established using an automatic routine. By combining the information from ECOCLIMAP and Hydro1k, the program avoids creating water divides within lakes. Land cover, soil information and topographic information (elevation and slope) from the grid cells included in each delineated sub-basin are also summarized in this process. Another example of systems used for automatic generation of the model input data from topographic databases is the TOPAZ system (Garbrecht & Martz, 1997).

Forcing data such as daily temperature and precipitation values for each simulated sub-basin are based on gridded global data sets such as ERA-40 (Uppala *et al*., 2006) or ERA-interim. These databases are derived from meteorological forecast model results which are re-analysed based on observed data assimilation. The WHIST is also capable of utilizing forcing data from climate projections (i.e. global and regional climate model results) to test various scenarios related to the impact of climate change on future water resources. For any of the databases, it is possible to use WHIST to complement the global data sets with local or regional information if such data are available and deemed to improve the accuracy of the model.

HYPE modelling

The HYPE model is a process-based, semi-distributed dynamic model, which integrates landscape elements and hydrological compartments along the flow paths with nutrient turnover and transport. Calculations are made on a daily time step in coupled sub-basins. Each sub-basin is divided into classes according to soil type, vegetation and altitude. The soil profile may be further divided into three layers. Model parameters are either general or related to soil type, or land cover. The model simulates e.g. snowmelt, surface runoff, surface erosion, macropore flow, tile drainage, groundwater outflow from the individual soil layers, nutrient turnover in soil, and transport/transformation in rivers and lakes. The model also accommodates a river routing routine, which enables the calculation of water discharge and nutrient transport through the mouth of each sub-basin. The

routine introduces lag times for water flowing in local streams and in the main rivers and mixing of local runoff water with water from upstream basins.

The HYPE model was applied on the La Plata Basin (LPB) in South America and continental Europe by using the WHIST system and global databases. For the European application these data sets were complemented with additional information on soil depth from the European Soils Database (ESDB). In addition, ERA-40 was replaced by ERAMESAN (Jansson *et al*., 2007): a gridded meteorological data set covering most of Europe on a 0.1º spatial resolution.

La Plata Basin is the fifth largest basin in the world, covering 3.6 million km^2 and extending over five countries. There is a large inter-basin hydroclimatic variation and the basin is very important for both hydroelectricity generation and agriculture (Mechoso *et al*., 2001). Calculations were done for approximately 4000 sub-basins. In the model set-up 11 monitoring sites were used for calibration of the LPB model and 10 independent sites were used for validation. The calibration methodology is described in detail in Donnelly *et al*. (2009). Many of the GRDC stations in the database only had data for short time periods. Stations with less than 10 years of data were excluded from the study. Additionally, stations where the upstream areas given by Hydro1k differed from the basin areas stated in the GRDC database by more than 25% were excluded. This was the case for 31% of the stations; for the remaining basins with area discrepancies, modelled streamflow was multiplied by a correction factor. Stations in the western Andean part of the basin were also excluded from the calibration data set, as it was seen early on that precipitation was highly overestimated in the ERA-40 data set for this part of the basin. The simulation period ranged from 1970 to 2000.

The European application covers about 7 million km^2 with calculations in some 8500 sub-basins, which gives an average resolution of about 1000 km^2. The calibration exercise for Europe is described in detail by Donnelly *et al*. (2009). In this paper, a validation of the model performance is carried out in the Baltic Sea drainage basin (Northern Europe), using 20 independent stations compared to the 32 stations used in the calibration of this region. The simulation period ranged from 1980 to 2000. In both studies, the relative volume error (VE) and the Nash-Sutcliffe model efficiency (NSE, Nash & Sutcliffe, 1970) were used as calibration criteria. The evaluation was carried out on daily streamflow values in Europe and on monthly streamflow values in the LPB.

A sensitivity analysis was also carried out to investigate the sensitivity of the model output to changes in key input data sets. The study was performed for the Daugava basin with a basin area of 87 900 km^2 draining to the Baltic Sea through its outlet in the Riga Bay. The sensitivity analysis included the following simulations: (a) 20% higher and lower precipitation: (b) 2ºC higher and lower temperature; (c) change of land cover in the whole basin to only forest and only open land; and (d) change of soil type in the whole basin to only clayey and only loamy soils.

RESULTS AND DISCUSSION

A majority of the calibration basins in Europe has NSE values greater than zero (Fig. 1), which is an indication that the model has some predictive power for streamflow variation (median for all sites = 0.27). The volume errors at the gauging stations in Europe (Fig. 1) indicate that the model performs well for many monitoring sites, showing volume errors of <10%. However, there are quite large differences in the results between regions that, to some extent, can be attributed to the input data sets and also to processes not yet included in the model. Few outliers are found in the northern and eastern parts of the area. Less agreement between modelled and observed water discharge is found in Central Europe, mainly linked to mountainous areas. This is partly due to the low accuracy of ERAMESAN in mountainous areas (Jansson *et al*., 2007). Moreover, there is a general underestimation in ECOCLIMAP of the glacial areas in the Alps. Other sources of error may be water abstraction and irrigation schemes, which are not yet accurately described in the model, but affect basin hydrology by subtracting water from streams and groundwater.

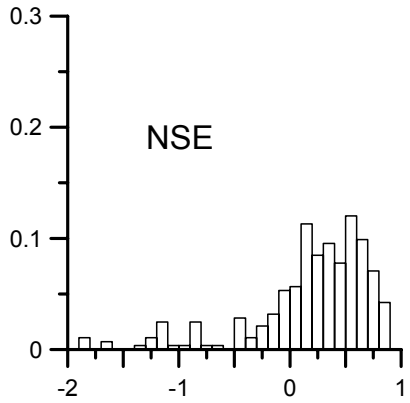

Fig. 1 Histogram of Nash-Sutcliffe model efficiency (NSE) values for the calibration sites in Europe: 6% of the sites have NSE < –2.

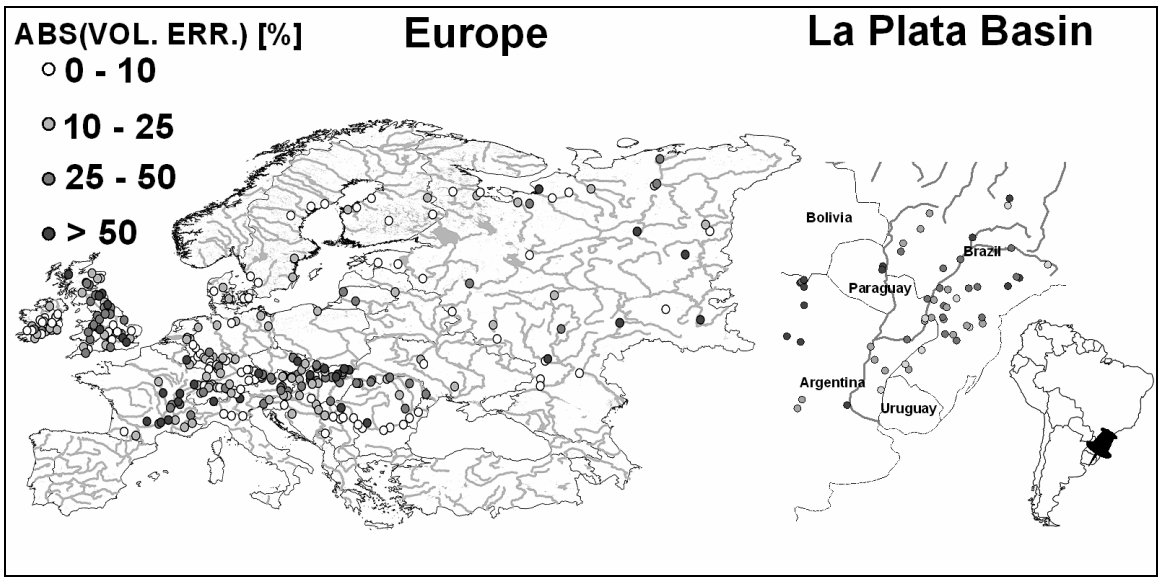

Fig. 2 Volume error in simulated river discharge at the calibration station for the two studied regions.

La Plata Basin shows large errors in simulated volume in the western part (Fig. 2). This is attributed to strong precipitation signal from the ERA-40 data over the Andean region resulting in an overestimation of runoff in these areas. This phenomenon has previously been recognised in modelling exercises over this region (Su & Lettenmaier, 2009). The effect is likely due to orographic enhancement of precipitation by the model underlying the ERA-40 data set. In reality, there is a well known interdecadal shift in weather patterns in the region (Berbery & Barros, 2001), which is not captured by the ERA-40 data.

In the rest of LPB, the volume error is generally <25% (median 2.5% for calibrated sites). This is a reasonable result, considering that the model set-up does not yet include large confined aquifers, wetlands, dams and regulations, which are significant in the modelled region. In the near future the global databases used for LPB-HYPE will be complemented with local meteorological data, more frequently monitored time-series of water discharge in several sites, river morphology, dams and regulation strategies of water power stations. Inclusion of these factors will affect the simulation and will provide more insight into the factors that are most important for improving the model performance.

The median NSE was 0.49 when simultaneous calibration was performed over 32 sites covering the Baltic Sea Basin (Table 2), indicating that the simulation results include more

Table 2 Selected model performance statistics for calibration and validation for the two study regions in the Baltic Sea Basin (Northern Europe) and the La Plata Basin (South America).

Region		No. of stations	NSE Median	NSE Max.	Volume error (%): Mean	Volume error (%): Median	Volume error (%): Std. dev.
Baltic Sea Basin	calibration	32	0.49	0.79	0.0	–1.4	14.8
	validation	20	0.21	0.77	16.1	2.5	37.0
La Plata Basin	calibration	11	–0.31	0.11	–3.4	–2.5	28.4
	validation	10	–0.96	–0.22	–16.7	–15.9	23.7

information about flow variation compared to the observed mean. For LPB, however, modelling did seldom add more information and maximum NSE was 0.11.

The model validation in independent sites (proxy-basin approach of Klemes, 1986) in the Baltic Sea Basin indicates that the model can be rather trustworthy for water balance estimation (Table 2), as the median VE is increased only slightly for the independent sites. However, four of the validation sites showed large over-estimations (>50%) of discharge, which explains the increased standard deviation. For the Baltic Sea Basin the model clearly gave added information about flow dynamics in ungauged basins, since the median NSE for independent sites was 0.21. However, these results are still low compared to what is found for modelling where data is more frequent and of good quality (c.f. Donnelly, 2009). It should be pointed out that some of the observed time series in the validation data set show clear signs of being heavily regulated. Lakes and dams are modelled using general rating curves in this application. This means that modelled hydrographs for regulated water courses will show a poor fit to the measurements.

It has already been mentioned that many of the available streamflow stations had to be removed from the analysis, as the basin delineation from Hydro1k did not agree with the real basin boundaries. This problem generally decreases with the size of the basin and new data sets such as HydroSHEDS (Lehner *et al*., 2008) may improve the situation. However, this must be considered as an extra source of error when modelling ungauged basins for which information on the real basin extent may not be known.

The simple sensitivity study shows that for Northern Europe the model results were most sensitive to forcing data and land cover (Fig. 3). It must be remembered that the results are

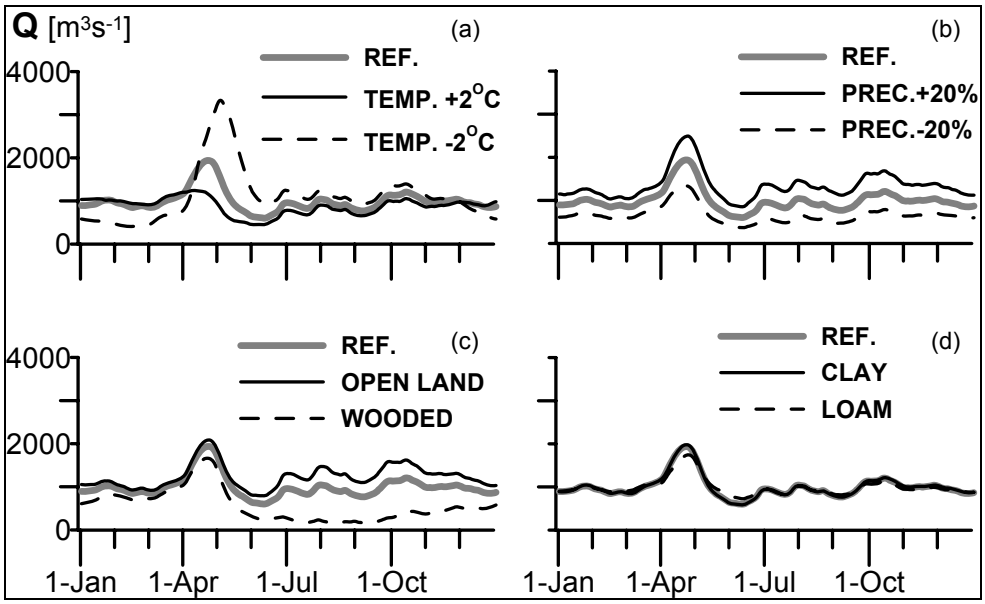

Fig. 3 Effect on modelled streamflow in the Daugava basin from change in (a) temperature, (b) precipitation, (c) land use, and (d) soil type. The figures show average values for each day of the year for the period 1980–2000.

influenced by the selected model structure and its parameterization (using different land covers and soil types). As expected, changing temperature has a large influence on the test basin (Fig. 3(a)). The spring flood is delayed and more pronounced when the temperature is lowered by 2°C. Increasing temperature resulted in a lower spring flood and lower flows during summer and early autumn due to increased evapotranspiration. The model is thus capable of responding to relative changes in climate and meteorological data, indicating that these data sets must be of a consistent spatial quality to allow for efficient modelling of ungauged basins.

The effect on runoff from a change in precipitation is large and relatively linear in this basin (Fig. 3(b)). A similar conclusion on the sensitivity of the model to errors in precipitation data was also drawn in La Plata Basin. The effect of changing land cover is relatively large for the Daugava basin model (Fig. 3(c)). Differences in simulated river discharge in the two land cover change scenarios tested is mainly due to differences in parameter values controlling the evapotranspiration rate between land with forest and open land. It is likely that the difference between the two land cover classes was exaggerated in the calibration exercise. It may, however, be an indication that a change to a more detailed database with higher resolution land cover information could improve model performance.

Changing soil type in the basin has little effect on modelled streamflow (Fig. 3(d)). Clayey soils were predominant in the basin, which explains why little change is seen in the model run using clay soils compared to the reference run. Changing soil type to a less hydrologically responsive loamy soil resulted in a more attenuated spring flood. Some parameters in the model are linked to soil type, but, in this model application, these were difficult to estimate due to few monitoring stations representing very different average soil conditions. More calibration sites may give other parameter values. Hence, it cannot be excluded that better soil information is still important for modelling water discharge.

CONCLUSIONS

A newly developed easily applicable system (WHIST) was used to generate model input data for hydrological multi-basin applications by compiling data from global data sets.

The results of the study indicate that the use of global data sets in hydrological modelling may be a valid concept for predictions in ungauged basins in data sparse regions. However, there are regional differences in the quality of the global data sets and limitations were found in all of the data sets examined.

Some areas are inherently more difficult to model than others due, for example, to complex geo-hydrological conditions or water management or regulations. It is not always clear whether poor model performance is caused by limitations in the input data or in the representation of hydrological processes in the model, or a combination of both. Further studies are required to better distinguish between specific sources of errors and possibilities for improvements of both databases and models.

Acknowledgements The authors would like to thank the Global Runoff Data Centre (GRDC) and the Baltic Sea Experiment (BALTEX) for providing streamflow gauging data.

REFERENCES

Arheimer, B., Lindström, G., Pers, C., Rosberg, J. & Strömqvist, J. (2008) Development and test of a new Swedish water quality model for small-scale and large-scale applications. In: *XXV Nordic Hydrological Conference* (Reykjavik, 11–13 August 2008), 483–492. *NHP Report no. 50*..

Berbery, E. H. & Barros, V. R. (2001) The hydrological cycle of the La Plata basin in South America. *J. Hydromet.* **3**, 630–645.

Champeaux, J. L., Masson, V. & Chauvin, F. (2005) ECOCLIMAP: a global database of land surface parameters at 1 km resolution. *Met. Appl*. **12**, 29–32.

Donnelly, C., Dahné, J., Lindström, G., Rosberg, J., Strömqvist, J., Pers, C., Yang, W. & Arheimer, B. (2009) An evaluation of multi-basin hydrological modelling for predictions in ungauged basins. In: *New Approaches to Hydrological Prediction in*

Data-sparse Regions (ed. by K. K. Yilmaz *et al.*) (Joint IAHS & IAH Convention, Hyderabad, India, 6–12 September 2009).. IAHS Publ. 333, this volume. IAHS Press, Wallingford, UK.

Garbrecht, J. & Martz, L. W. (1997) TOPAZ Version 1.20: An automated digital landscape analysis tool for topographic evaluation, drainage identification, watershed segmentation and subcatchment parameterization – Overview. Rep. #GRL 97-2, Grazinglands Research Laboratory, US Dept Agric., Agricultural Research Service, El Reno, Oklahoma, USA.

GRDC (2008) Global Runoff Data Centre, World Meteorological Organisation & The German Federal Institute of Hydrology. http://www.bafg.de/cln_007/nn_301072/GRDC/Home/homepage__node.html?__nnn=true. (Accessed 16/2-2009).

Jansson, A., Persson, C. & Strandberg, G. (2007) 2D meso-scale re-analysis of precipitation, temperature and wind over Europe – ERAMESAN time period 1980–2004. *SMHI reports: Meteorology and climatology no. 112.*

Klemes, V. (1986) Operational testing of hydrological simulation models. *Hydrol. Sci. J.* **31**, 13–24.

Lehner, B., Verdin, K. & Jarvis, A. (2008) New global hydrography derived from spaceborne elevation data. *Eos, Trans. Am. Geophys. Union* **89**(10), 93–94.

Lindström, G., Pers, C. Rosberg, J., Strömqvist, J. & Arheimer, B. (2009) Development and test of the HYPE (Hydrological Predictions for the Environment) model – A water quality model for different spatial scales. *Hydrol. Res.*, submitted 7/1 - 2009.

Mechoso, C. R., Dias, P. S., Baetghen, W., Barros, V., Berbery, E. H., Clarke, R., Cullen, H., Ereño, C., Grassi, B. & Lettenmaier, D. (2001) Climatology and Hydrology of the Plata Basin. VAMOS/CLIVAR document, http://www.clivar.org/organization/vamos/publications/laplata.pdf (accessed 16 February 2009).

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

Su, F. & Lettenmaier, D. P. (2009) Estimation of surface water budget of La Plata Basin. *J. Hydromet.* In press.

Uppala, S. M., Kållberg, P. W., Simmons, A. J., Andrae, U, da Costa Bechtold, V., Fiorino M., Gibson, J. K., Haseler, J., Hernandez, A., Kelly, G. A., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R. P., Andersson, E., Arpe, K., Balmaseda, M. A., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins, B. J., Isaksen, L., Janssen, P. A. E. M., McNally, A. P., Mahfouf, J.-F., Jenne, R., Morcrette, J.-J., Rayner, N. A., Saunders, R. W., Simon, P., Sterl, A., Trenberth, K. E., Untch, A., Vasiljevic, D., Viterbo, P. & Woollen, J. (2006) The ERA-40 reanalysis. *Quart. J. Roy. Met. Soc.* **131**, 2961–3012.

USGS (US Geological Survey) (2000) Hydro1k Elevation Derivative Database. http://edc.usgs.gov/products/elevation/gtopo30/hydro/index.html. (accessed 25 February 2009).