Dr. Lukas Gudmundsson
Institute for Atmospheric and
Climate Science ETH Zurich
Universitaetsstrasse 16
CH-8092 Zurich
Phone: +41 44 632 77 09
E-mail: lukas.gudmundsson@env.ethz.ch

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

HESSD Editorial Team

Zürich, 5 May 2015

Re-submission of manuscript entitled "Towards observation based gridded runoff estimates for Europe" to *HESSD*

Dear Mr. Andréassian,

We thank you for forwarding us the comments of referee #3 and for offering us the opportunity to revise the manuscript. Based on the referee's comments following additions to the manuscript were made:

(1) On request of referee #3 we expanded the paragraph that introduces the idea to use machine learning to estimate runoff at ungauged locations.
(2) Following the suggestion of referee #3 we restructured the data section, now presenting the LSM data as an independent data category ("2.3 Comparison Data")
(3) As requested by the referee section 4.2.3 now provides more details on the inconsistency between observed streamflow and gridded precipitation in Scandinavia.

Detailed answers to the referee's suggestions are provided in the attached document. We are confident that these additions help improve the readability of the manuscript.

Yours sincerely,

Lukas Gudmundsson

Lukas Gudmundsson

Response to Reviewer #3

Lukas Gudmundsson & Sonia I. Seneviratne

May 5, 2015

We would like to thank Reviewer #3 for reporting his/her constructive suggestions. In the following we provide point by point answers to his/her comments. For the sake of clarity we first repeat the reviewer's comments (*in italic*) and then provide our response.

I was not involved in the first round of revision, and I thus take this revised paper as a new one.

SYNTHESIS
1. In simple words, I would tell that this paper presents a method to upscale and then interpolate the observed streamflow from small catchments. A validation is proposed based on (i) cross validation, (ii) large catchments which were not part of the initial catchment dataset, (iii) independent evaporation measurements. A comparison is provided based on a combination of LSM data.

We appreciate this concise summary of our research.

WHAT I LIKED IN THIS PAPER
2. This paper is scientifically sound (I appreciated the three steps validation step in particular). It is rather well written (although sometimes I feel that the authors use unneeded complex words for things which are after all very simple).

We would like to thank the referee for this positive evaluation of our research.

WHAT I DID NOT LIKE SO MUCH IN THIS PAPER
*3. I always have problem with machine learning techniques which I like to classify as "fancy statistical methods" because they are black-box (sometimes only grey...) sometimes difficult to follow and always difficult to INTERPRET PHYSICALLY. My preferred approach would have been to use a simple regression (Q=a*P^b*E^c... eventually with monthly varying parameters, because these parameters are easy to interpret) and THEN use the machine learning technique to interpolate the residual. It would probably have given exactly the same efficiency but it would have been much easier to interpret. I have no objections to have a black-box approach to map*

residuals (because by definition, the residual is impossible to understand), I always consider it a pity to mix what is understandable with what is not.

We fully agree with the referee on the merits of simple regression models, which can help to disentangle contributions of different predictors on a target variable. However, the primary aim of the presented study is not to focus on physical understanding but on developing a framework that can be used to estimate runoff at ungauged locations. To this end we decided to follow the approach of Jung et al. (2009, 2010, 2011) who used machine learning techniques to estimate land-atmosphere fluxes at the global scale. The advantage of this approach is that it does not rely on strong prior assumptions and hence reduces the danger of biases that are related to model formulation. Consequently, we expect machine learning techniques also to capture relevant processes which influence the data, but might not be considered in a parametric (e.g. additive or multiplicative) model setup. This is also nicely illustrated by Beck et al. (2013, Figure 6). In the present context, the choice of machine learning over parametric regression is consequently a tradeoff between predictive power and the possibility to interpret the model structure. As the aim of this study is to produce reliable estimates of European runoff we opted for increased predictive power.

To clarify this aspect to the reader we extended the paragraph, introducing the approach of Jung et al. (2009, 2010, 2011) accordingly.

RECOMMANDATIONS

4. You must do something about the negative RFM values in Fig. 11 you cant just write that you know they exist!

See our next answer.

5. Fig 11 : there is certainly a lot of information to gather from the differences in Fig 11. May I suggest mapping the differences in the Turc-Budyko non dimensional graph (i.e. x=P/E0 and y=Q/P, each catchment is represented by a point and the colour of the point would present the difference between the two models), this could help explain physically the differences.

Reviewer #3 correctly notes that that the longterm mean difference between precipitation and estimated runoff is negative in Scandinavia, pointing towards physically inconsistent data. This inconsistency does already emerge at small scales: For a large portion of Scandinavian grid-cells, the runoff coefficients Q/P are larger than one, suggesting that the WATCH forcing data underestimate precipitation in this region. This issue has been extensively documented in previous studies (Gudmundsson et al., 2012; Kauffeldt et al., 2013). For convenience Figure A1 provides a reproduction of Figure 1c in Gudmundsson et al. (2012), which systematically documents this issue. Kauffeldt et al. (2013) conclude that these differences mostly occur in regions where snow undercatch is an issue.

The statistical runoff model accounts for this bias and consequently the difference between WFD precipitation and estimated runoff is negative. This artefact is an intrinsic feature of the data. Therefore we refrain from further modifications such as ad-hoc adjustments of precipitation in Scandinavia as they would incorporate a subjective component into our analysis. Only an un-biased precipitation estimator would resolve this issue.

We expanded the section discussing this results aiming at communicating the source of this bias more systematically to the reader. We did however refrain from adding additional quantitative results to the paper as the inconsistency between runoff observations and WFD precipitation is already documented in the literature (Gudmundsson et al., 2012; Kauffeldt et al., 2013).

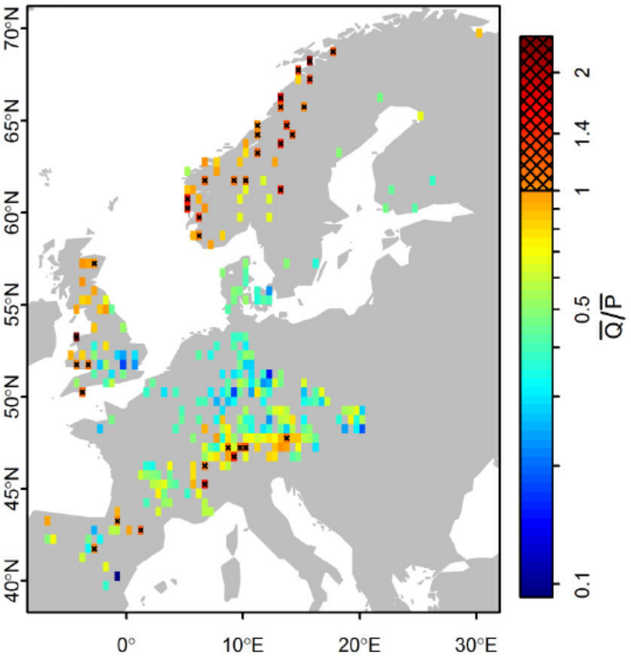

Figure A1: Long-term runoff coefficient ($\bar{Q}/\bar{P}$) derived from observed runoff ($\bar{Q}$) and WFD precipitation ($\bar{P}$). Shaded areas indicate $\bar{Q}/\bar{P} \geq 1$. Reproduction of Figure 1c in Gudmundsson et al. (2012).

CONCLUSION

Overall, I liked this paper and I recommend its publication. Here are my suggestions to improve its readability:

6. You should try to synthesize your approach using simpler words in the introduction (for example, what you do is a spatial interpolation but I did not find the word in your paper);

We thank Referee #3 for this suggestion. To account for this, we did extend the paragraph introducing the approach of Jung et al. (2009, 2010, 2011), emphasising that it is basically a (non-linear) regression.

Note also, that we do not do an "spatial interpolation" as suggested by the referee. We rather build a non-parametric regression model that maps monthly runoff as a function of atmospheric conditions. This would correspond to an "interpolation in the atmospheric variable space". As this could be easily confused with spatial interpolation we prefer not to use the term "interpolation" in the article.

7. Line 122 : you should not include LSM runoff in the validation data paragraph... it is just a comparison (just change : 2.2 Comparison data and then 2.3 Validation data

We thank Referee #3 for this suggestion. The sectioning was changed accordingly.

8. Line 324 : do something for those negative values.

See our corresponding reply above.

References

Beck, H. E., van Dijk, A. I. J. M., Miralles, D. G., de Jeu, R. A. M., (Sampurno) Bruijnzeel, L. A., McVicar, T. R., and Schellekens, J.: Global patterns in base flow index and recession based on streamflow observations from 3394 catchments, Water Resources Research, 49, 7843–7863, doi:10.1002/2013WR013918, 2013.

Gudmundsson, L., Wagener, T., Tallaksen, L. M., and Engeland, K.: Evaluation of nine large-scale hydrological models with respect to the seasonal runoff climatology in Europe, Water Resour. Res., 48, W11 504, doi:10.1029/2011WR010911, 2012.

Jung, M., Reichstein, M., and Bondeau, A.: Towards global empirical upscaling of FLUXNET eddy covariance observations: validation of a model tree ensemble approach using a biosphere model, Biogeosciences, 6, 2001–2013, doi:10.5194/bg-6-2001-2009, 2009.

Jung, M., Reichstein, M., Ciais, P., Seneviratne, S. I., Sheffield, J., Goulden, M. L., Bonan, G., Cescatti, A., Chen, J., de Jeu, R., Dolman, A. J., Eugster, W., Gerten, D., Gianelle, D., Gobron, N., Heinke, J., Kimball, J., Law, B. E., Montagnani, L., Mu, Q., Mueller, B., Oleson, K., Papale, D., Richardson, A. D., Roupsard, O., Running, S., Tomelleri, E., Viovy, N., Weber, U., Williams, C., Wood, E., Zaehle, S., and Zhang, K.: Recent decline in the global land evapotranspiration trend due to limited moisture supply, Nature, 467, 951 – 954, doi:10.1038/nature09396, 2010.

Jung, M., Reichstein, M., Margolis, H. A., Cescatti, A., Richardson, A. D., Arain, M. A., Arneth, A., Bernhofer, C., Bonal, D., Chen, J., Gianelle, D., Gobron, N., Kiely, G., Kutsch, W., Lasslop, G., Law, B. E., Lindroth, A., Merbold, L., Montagnani, L., Moors, E. J., Papale, D., Sottocornola, M., Vaccari, F., and Williams, C.: Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations, J. Geophys. Res., 116, G00J07, doi:10.1029/2010JG001566, 2011.

Kauffeldt, A., Halldin, S., Rodhe, A., Xu, C.-Y., and Westerberg, I. K.: Disinformative data in large-scale hydrological modelling, Hydrology and Earth System Sciences, 17, 2845–2857, doi:10.5194/hess-17-2845-2013, 2013.

Manuscript prepared for Hydrol. Earth Syst. Sci.
with version 5.0 of the LaTeX class copernicus.cls.
Date: 5 May 2015

Towards observation based gridded runoff estimates for Europe

Lukas Gudmundsson and Sonia I. Seneviratne

Institute for Atmospheric and Climate Science, ETH Zurich, Universitaetstrasse 16, 8092 Zurich, Switzerland

Correspondence to: Lukas Gudmundsson & Sonia I. Seneviratne
(lukas.gudmundsson@env.ethz.ch & sonia.seneviratne@ethz.ch)

Abstract. Terrestrial water variables are the key to understanding ecosystem processes, feed back on weather and climate, and are a prerequisite for human activities. To provide context for local investigations and to better understand phenomena that only emerge at large spatial scales, reliable information on continental scale freshwater dynamics is necessary. To date streamflow is among the best observed variables of terrestrial water systems. However, observation networks have a limited station density and often incomplete temporal coverage, limiting investigations to locations and times with observations. This paper presents a methodology to estimate continental scale runoff on a 0.5° spatial grid with monthly resolution. The methodology is based on statistical up-scaling of observed streamflow from small catchments in Europe and exploits readily available gridded atmospheric forcing data combined with the capability of machine learning techniques. The resulting runoff estimates are validated against (1) runoff from small catchments that were not used for model training, (2) river discharge from nine continental scale river basins and (3) independent estimates of long-term mean evapotranspiration at the pan-European scale. In addition it is shown that the produced gridded runoff compares on average better to observations than a multi-model ensemble of comprehensive Land Surface Models (LSMs), making it an ideal candidate for model evaluation and model development. In particular, the presented machine learning approach may help determining which factors are most relevant for an efficient modelling of runoff at regional scales. Finally, the resulting data product is used to derive a comprehensive runoff-climatology for Europe and its potential for drought monitoring is illustrated.

1 Introduction

Terrestrial water storages and fluxes are key variables in the Earth system, as they are a primary control for many ecosystem processes (e.g. Ciais et al., 2005; Granier et al., 2007; Reichstein et al., 2013; Guan et al., 2015), influence weather and climate through land-atmosphere interactions (e.g. Koster et al., 2004; Seneviratne et al., 2010) and are the basis for many human activities (e.g Döll et al., 2009; Vörösmarty et al., 2010; Orlowsky et al., 2014). Consequently information of the historical space and time evolution of variables such as evapotranspiration, soil moisture, groundwater and runoff are of great interest. However, most of these variables are only observed at few locations in space and often with irregular temporal coverage, limiting analysis to the well monitored regions. Consequently data products providing reliable estimates of the historical space-time evolution of these variables for large, continental scale regions are of vital importance. Such data products will not only allow to investigate terrestrial water dynamics at locations without observations, but more importantly also allow the study of processes and phenomena that emerge on large, continental, scales. Such studies include but are not limited to: (1) The analysis of fresh water climatologies (e.g. Dettinger and Diaz, 2000; Fekete et al., 2002; Reager and Famiglietti, 2013); (2) The assessment of large-scale droughts (e.g. Sheffield et al., 2012; Tallaksen and Stahl, 2014; Thomas et al., 2014; Gudmundsson et al., 2014; Gudmundsson and Seneviratne, 2015); (3) The validation of Land Surface Models (LSMs) and hydrological models used at large scales (e.g. Dirmeyer et al., 2006; Haddeland et al., 2011; Gudmundsson et al., 2012a,b; Schewe et al., 2014); (4) Investigating the link between climate variability and terrestrial water dynamics, including feedbacks (e.g. Tootle and Piechota, 2006; Jung et al., 2010; Gudmundsson et al., 2011b; Mueller and Seneviratne, 2012; de Linage et al., 2014; Miralles et al., 2014); and (5) Analysing the effect of climate change on freshwater resources (e.g. Krakauer and Fung, 2008; Stahl et al., 2012; Famiglietti and Rodell, 2013; Greve et al., 2014).

To date, two main approaches for continental to global scale estimation of terrestrial water dynamics are in use. The first approach is based on LSMs that are driven by historical atmospheric forcing (e.g. Fekete et al., 2002; Rodell et al., 2004; Dirmeyer et al., 2006; Fekete et al., 2011; Balsamo et al., 2013). While LSM-based estimates are attractive because they provide comprehensive information on a large number of relevant variables, the resulting data are still highly model dependent and large uncertainties remain (e.g Haddeland et al., 2012; Gudmundsson et al., 2012a,b; Mueller et al., 2011b, 2013; Prudhomme et al., 2014). In recent years, the rapid evolution of satellite remote sensing has allowed to provide estimates of selected variables including soil moisture (e.g. Wagner et al., 2007; de Jeu et al., 2008; Seneviratne et al., 2010) and total terrestrial water storage (e.g Houborg et al., 2012; Landerer and Swenson, 2012; Rodell and Famiglietti, 1999; Famiglietti and Rodell, 2013). However, satellite observations only cover a relatively short time window and issues such as inhomogeneities due to changes in ~~instrumentations~~ instrumentation and uncertainties in retrieval algorithms are still limiting their application (Loew et al., 2013; Hirschi et al., 2014).

A common feature of the above mentioned approaches is that they only exploit in-situ observations of terrestrial water variables to a very limited degree. Historically, catchment runoff is likely the best monitored variable of terrestrial water systems, which has been observed for centuries to decades at thousands of locations covering the entire globe ~~(Slack and Landwehr, 1992; Hannah et al., 2011)~~ (Slack and Landwehr, 1992; Hannah et al., 2011; Fekete et al., 2012) . Other variables such as evapotranspiration or soil moisture have received less attention and consequently respective ground observations are available at fewer locations and often cover much shorter time periods (Baldocchi, 2008; Seneviratne et al., 2010; Dorigo et al., 2013). Nevertheless ~~recent studies (Jung et al., 2009, 2010, 2011) succeeded to up-scale in-situ observations~~ Jung et al. (2009, 2010, 2011) successfully derive global estimates of evapotranspiration, sensible heat flux and carbon exchange ~~to regular spatial grids using machine learning techniques. Combining the quality of~~ on the basis of in-situ observations of the FLUXNET observatories (Baldocchi, 2008) ~~, the availability of gridded explanatory variables and the versatility of modern machine learning they derived global estimates of evaportranspiration and carbon fluxes with monthly resolution on regular spatial grids.~~ using machine learning techniques. Technically speaking, Jung et al. (2009, 2010, 2011) did build a nonlinear regression model that predicts land-atmosphere fluxes as a function of gridded atmospheric variables (e.g. precipitation) and remotely sensed information on vegetation activity. As the nonlinearity of the underlying processes prevents the identification of parametric regression models, the application of machine learning is necessary. This is also illustrated by Beck et al. (2013) , who used neural networks to estimate global maps of several streamflow characteristics, including the base flow index.

~~This study suggests~~ The presented study proposes a framework for estimating the historical space-time evolution of runoff in Europe on the basis of observations from small catchments. Following Jung et al. (2009, 2010, 2011) we combine the advantage of in situ observations and the availability of gridded atmospheric observations with machine learning techniques to derive estimates of monthly runoff in Europe on a regular spatial grid. The accuracy of the estimated runoff fields is assessed with respect to data that were not used for model identification and compared to an ensemble of comprehensive land surface models. Finally, example applications of the resulting data product are provided and implications from the empirical modelling exercise are discussed in the context of physical model development.

2 Data

2.1 Modelling data

2.1.1 Atmospheric forcing

Estimates of atmospheric near-surface variables were taken from the WATCH Forcing Data (WFD, Weedon et al., 2011) which are available on a regular $0.5^\circ \times 0.5^\circ$ grid. The WFD were developed in the context of the WATCH (Water and Global Change) project (http://www.eu-watch.org/, accessed: June 24 2014). The analysis is based on the full WFD, covering the following set of variables: Rainfall, snowfall, air temperature, incoming long and short wave radiations, humidity, surface pressure and wind speed. The WFD are available at sub-daily resolution and were aggregated to monthly mean values.

2.1.2 Runoff observations

The investigation is based on 426 streamflow series from small undisturbed catchments, covering the 1963 - 2000 time period (Figure 1). The data are a subset (see Stahl et al., 2010, for details) of the European Water Archive (EWA). The EWA is collected by the European Flow Regimes from International Experimental and Network Data (Euro-FRIEND) project (http://ne-friend.bafg.de/servlet/is/7413/, accessed: June 24 2014) and held by the Global Runoff Data Centre (GRDC, http://grdc.bafg.de, accessed: June 24 2014).

As the majority of the considered catchments is much smaller than the 0.5° grid cells of the atmospheric forcing data (Figure 1), the time series of the individual catchments were assigned to the corresponding grid cells. Following previous studies (Arnell, 1995; Gudmundsson et al., 2011b, 2012b,a), streamflow observations from the individual catchments were first converted into runoff rates per unit area and the coordinates of the gauging stations were assigned to the 0.5° grid cells defined by the atmospheric forcing data. If more than one gauging station occurred in one catchment, the catchment area weighted average runoff rate was used. This procedure results in 298 grid cells with observed daily runoff rates, which were subsequently aggregated to mean monthly values (Figure 1).

In the following the monthly mean grid-cell averaged runoff rates are referred to as "observed runoff". Although streamflow, which is used to compute these estimate, is different from runoff we argue that the differences between the two quantities become small at the considered space and time scales. The main difference between streamflow and runoff is that the former has been routed trough a channel network. However, the associated processes operate on time scales that are much smaller than the resolution of the presented analysis. For example, hydrograph wave speeds are approximately $0.5\,\mathrm{m}\ \mathrm{sec}^{-1} = 1.8\,\mathrm{km}\ \mathrm{h}^{-1} = 43.2\,\mathrm{km}\ \mathrm{day}^{-1}$ (e.g. Wong and Laurenson, 1983), implying that at least daily resolution would be required to resolve these processes for 0.5° grid-cells. At monthly time scales, however, total catchment runoff can be assumed to equal the sum of streamflow, if water losses through e.g. channel evaporation are negligible. As the presented study operates on monthly resolution and on a 0.5° grid ($\approx 50\,\mathrm{km}$), it is consequently unlikely that effects of channel routing will impair the results.

2.1.3 Land parameters

Median grid-cell slope was derived from the HYDRO1k dataset which is available from the U.S. Geological Survey (Figure 2). Information on soil texture for each ~~grid-cell~~ grid cell (median fraction of clay, silt, sand, gravel) were taken from the Harmonized World Soil Database (version 1.2) (FAO et al., 2012) (Figure 3).

2.2 Validation data

2.2.1 ~~LSM runoff~~

~~The results of the statistical modelling exercise were also compared to runoff simulations from nine state-of-the-art LSMs, developed by the WATCH project. Details on the simulation setup, key features of the participating models, and further model validation can be found in the literature (Haddeland et al., 2011; Gudmundsson et al., 2012b,a) . All participating models were forced using the WFD which guarantees a fair comparison with the statistical runoff estimates introduced in this study. The LSM runoff simulations were augmented by the multi-model mean (MMM).~~

2.2.1 Continental scale river discharge

Observed monthly discharge from nine continental scale river basins (Ebro, Elbe, Garonne, Loire, Po, Rhine, Rhone, Seine, Weser) and corresponding catchment shapes where taken from a previously assembled collection (see Hirschi et al. (2006) and Mueller et al. (2011a) for details).

2.2.2 Long-Term mean evapotranspiration

A comprehensive estimate of the long-term mean (1989 - 1995) land evapotranspiration was taken from the LandFlux-EVAL synthesis product (Mueller et al., 2013), which combines informations from 40 distinct evapotranspiration estimates on a 2° grid.

2.3 Comparison Data

The results of the statistical modelling exercise were also compared to runoff simulations from nine state-of-the-art LSMs, developed by the WATCH project. Details on the simulation setup, key features of the participating models, and further model validation can be found in the literature (Haddeland et al., 2011; Gudmundsson et al., 2012b,a) . All participating models were forced using the WFD which guarantees a fair comparison with the statistical runoff estimates introduced in this study. The LSM runoff simulations were augmented by the multi-model mean (MMM).

3 Methods

3.1 Statistical model setup

The aim of this study is to estimate monthly runoff, $Q_{x,t}$, at different land units x and time steps t. To achieve this, $Q_{x,t}$ is related to a set of explanatory variables that are available at all locations within the spatial domain through a machine learning model h, which is described in detail in Section 3.2.

We derive three models, of various degrees of complexity. The simplest case assessed in this study is solely based on gridded precipitation, $P_{x,t}$, and temperature $T_{x,t}$ such that

$$Q_{x,t} = h(\tau_n(P_{x,t}), \tau_n(T_{x,t})), \tag{1}$$

where the time lag operator τ_n is defined as $\tau_n(X_{x,t}) = [X_{x,t}, X_{x,t-1}, \dots, X_{x,t-n}]$ and gives access to atmospheric conditions over the past n time steps (months). This time lag operator allows to approximate storage effects that are relevant for runoff generation. In the presented analysis, input from the previous year is considered ($n = 11$), which enables the model to take limited storage processes related e.g. to groundwater and snow into account. Note also that the model h is only identified once and applicable at all locations in space. This implies that all information on spatial variability only comes from the atmospheric input data. As the WFD provides separate information on rain and snowfall, precipitation is here defined as the sum of both components. This simple setup is motivated by the tradition that runoff modelling at catchment scales relies in many cases only on precipitation and temperature forcing.

The second model setup is defined as

$$Q_{x,t} = h(\tau_n(I^1_{x,t}), \tau_n(I^2_{x,t}), \dots, \tau_n(I^p_{x,t})), \tag{2}$$

where $I^1_{x,t}, \dots, I^p_{x,t}$ are all atmospheric forcing variables available within the WFD (see Section 2.1.1). The rationale underlying this approach is that processes such as evapotranspiration and snow dynamics do not only depend on precipitation and temperature but also on many other forcing variables including humidity, wind speed and different radiation components.

Finally the most complex model setup is specified as

$$Q_{x,t} = h(\tau_n(I^1_{x,t}), \tau_n(I^2_{x,t}), \dots, \tau_n(I^p_{x,t}), \Pi_x), \tag{3}$$

where Π_x is a vector, containing information on slope and soil texture (see Section 2.1.1). The idea underlying this last setup is to increase the realism of the statistical model, as terrestrial water dynamics is not only dependent on atmospheric forcing but also on local variations in land properties which influence runoff generation.

3.2 Model identification

The practical challenge in the application of Equations 1 to 3 is the identification of the model h. For this we follow Jung et al. (2009, 2010, 2011) and exploit the capability of modern machine learning techniques. In contrast to Jung et al. (2009, 2010, 2011), who used Model Tree Ensembles, we employ here a closely related method called Random Forests (RF) (Breiman, 2001). The use of RF is a pragmatic choice, as this technique is well established, requires only few user specifications (see e.g. Hastie et al., 2009) and is implemented in standard software environments (e.g. Liaw and Wiener, 2002). Note, however, that other machine learning tools such as Boosting techniques, Neural Networks or Support Vector Machines are likely to have similar performance (e.g. Bishop, 2006; Hastie et al., 2009).

Technically, RF are based on large ensembles of a modified version of Classification and Regression Trees, each grown on a bootstrap sample of the data. Despite its considerable complexity, the RF algorithm (Breiman, 2001; Liaw and Wiener, 2002; Hastie et al., 2009) can be summarised in a simplified manner as:

1. Draw B bootstrap samples from the data.
2. For each bootstrap sample, grow a Random Forest tree by recursively repeating the following steps:
    (a) Select m of the available predictor variables at random.
    (b) Among the m selected variables: find the one with the split point that best partitions the data.
    (c) Split the data into two nodes and repeat the two previous steps on each node until the terminal node has reached the minimum node size n.
3. The RF prediction for new data is the average of the predictions of the B individual trees.

The free parameters of RFs need to be specified by the user. We opted for $B = 1000$, $n = 10$, and $m = p/3$, where p is the number of predictor variables, following recommendations in the literature (Hastie et al., 2009). In general, we found the results to be little sensitive to the parameter choice as long as the number of grown trees (B) was large enough.

3.3 Model selection and validation

3.3.1 Cross validation

An important issue in statistical modelling is the fact that using the same data for model identification and model evaluation can result in too optimistic estimates of model performance. Therefore, the results of machine learning tools are commonly assessed using K fold cross validation (e.g. Bishop, 2006; Hastie et al., 2009). Cross-validation guarantees that the data used for model validation are independent from the data used for model identification. For cross validation, the data are first randomly split into K subsamples. Subsequently one of the subsamples is removed and the model is trained on the remaining $K - 1$ subsamples. Finally the resulting model is used to predict the data that have been left out. These steps are repeated K times until each subsample has been left out once.

The procedure consequently results in predictions of the data that are independent of the data used for model identification.

To enhance the interpretability of cross validation in the context of this study we focus on the following two modifications of the usual cross validation procedure: In a first experiment, the focus is on the models ability to estimate runoff at spatial locations (x) that were not used for model identification. For this, the grid cells with observations are randomly split into $K = 10$ subsamples, which were successively left out for model training. This procedure guarantees that at each location with observations, model estimates are available that are independent of the data used for model identification. In the following we refer to this procedure as "cross validation in space". Note that this validation strategy makes the analysis compatible with the Prediction of Ungauged Basins (PUB) initiative (Sivapalan et al., 2003; Blöschl et al., 2013; Hrachowitz et al., 2013; Parajka et al., 2013) of the International Association of Hydrological Sciences (IAHS). In a second experiment the focus is on the models' ability to estimate runoff dynamics at time steps (t) that were not used for model identification. For this, the data were split into $K = 10$ continuous time blocks, which were successively left out once for model training. This procedure is referred to as "cross validation in time" and provides estimates of runoff at time steps that were not used for model identification.

3.3.2 Model Selection

Model selection is based on the total root mean square error, integrating model accuracy over space and time:

$$\text{RMSE} = \sqrt{\sum_{x,t}(m_{x,t} - o_{x,t})^2}, \tag{4}$$

where $m_{x,t}$ and $o_{x,t}$ refer to the modelled and observed values respectively. RMSE for each of the candidate models (Section 3.1) is estimated based on the two cross validation experiments. Uncertainty of the RMSE is quantified using 95% bootstrap confidence intervals with 2000 replications.

3.3.3 Model validation

Model performance is assessed for individual grid cells, where o_t refers to the observed and m_t to the modelled runoff series. Model performance is quantified using six different performance metrics, each focusing on different aspects of runoff dynamics:

1. The seasonal cycle skill score (Wilks, 2011) is defined as

$$S_{\text{seas}} = 1 - \frac{\sum_t (m_t - o_t)^2}{\sum_t (m_t - \text{seas}(o_t))^2}, \tag{5}$$

where $\text{seas}(o_t)$ refers to the long-term mean runoff for each month. $S_{\text{seas}} \, \epsilon \, (-\infty, 1]$ and positive values indicate that the model is on average closer to the observations than the mean annual cycle.

2. The model efficiency (Nash and Sutcliffe, 1970; Wilks, 2011) is defined as

$$\mathrm{MEf} = 1 - \frac{\sum_t (m_t - o_t)^2}{\sum_t (m_t - \mathrm{mean}(o_t))^2}, \tag{6}$$

where mean(o_t) refers to the long-term mean of the observation. $S_{\mathrm{MEf}} \epsilon (-\infty, 1]$ and positive values indicate that the model is on average closer to the observations than the mean of the observations.

3. The relative model bias is defined as

$$\mathrm{BIAS} = \frac{\mathrm{mean}(m_t - o_t)}{\mathrm{mean}(o_t)}, \tag{7}$$

i.e. the mean difference between observed and modelled values scaled by the mean of the observations. The optimal value is zero and positive (negative) values indicate overestimation (underestimation) of the mean runoff.

4. The coefficient of determination (squared correlation coefficient), R^2, measures the agreement between the temporal evolution of the modelled and observed series.

5. The coefficient of determination between the observed and the modelled mean annual cycle, Climatology-R^2, is sensitive to differences in the phasing of the mean annual cycle.

6. The coefficient of determination between the monthly anomalies (i.e. monthly time series with the long-term mean of each month removed), Anomaly-R^2, indicates the agreement between observed and modelled values after removing the mean seasonal cycle.

4 Results

4.1 Model Selection

Figure 4 shows the RMSE of the Random Forest Model (RFM) for all three model setups and both cross validation experiments. For the cross validation in space, the model that only depends on precipitation and temperature (Equation (1)) has the largest error and the two other models (Equations (2) and (3)) have almost equal performance. The situation differs for the cross validation in time. Here the model with full atmospheric forcing (Equation (2)) significantly outperforms the other two models. As the model with full atmospheric forcing shows the best performance in both cross validation experiments it was selected and is considered for further analysis. In the following RFM refers to this selected model, unless specified differently.

4.2 Model Validation

4.2.1 Grid-cell scale validation

Figure 5 shows the RMSE of the RFM, derived from the cross validation in space experiment at each grid cell with observations as well as time series of observed and modelled runoff at the grid cells with the smallest, the median and the largest error. The grid cell error shows some spatial patterns, with a tendency to increase in mountainous regions where observed runoff rates are highest. The selected time series allow for a qualitative assessment of the strengths and shortcomings of the RFM, indicating a good agreement of observed and modelled runoff, but also highlighting some deficiencies in capturing peak flows.

A more comprehensive overview on model performance is provided in Figures 6 and 7, which show the spatial distribution of all considered skill scores of the selected RFM for both the cross validation in time and for the cross validation in space. Table 1 lists the median performance for both cross-validation experiments. In addition the boxplots in Figures 6 and 7 allow to compare the distribution of the performance of all considered modelling setups (Equations (1) - (3)) to the performance of LSM simulations. For the sake of brevity the following description of the results is limited to the selected RFM with full atmospheric forcing. Overall, there are no clear spatial patterns in S_{seas} and MEf which are on average positive for both cross validation experiments. This shows that the RFM is at most locations a better estimator of monthly runoff variability than mere repetitions of the climatology. Interestingly the RFM also outperforms all LSMs under consideration with respect to S_{seas} and MEf. On average the relative BIAS of the RFM is slightly negative, indicating a tendency of the model to underestimate monthly runoff rates in the considered catchments. Generally the relative bias of the considered LSMs is comparable to the RFM bias highlighting their similar mean annual runoff rates. The median coefficient of determination, R^2, between the RFM and the observed runoff rates are high and there are no pronounced spatial patterns for both cross validation experiments. This indicates the capability of the empirical model to capture the temporal evolution of runoff in Europe. Also with respect to R^2, the selected RFM is closer to the observations than any LSM under consideration. The remarkably high coefficient of determination between the observed and modelled mean annual cycles, Climatology-R^2, of the RFM are contrasted by the relatively low correlations of the LSMs. This result highlights the RFM's ability to capture the seasonality of runoff, but also points towards the fact that the considered LSMs have issues with reproducing this feature. The median coefficient of determination of observed and modelled monthly runoff anomalies, Anomaly-R^2, reach only intermediate levels showing the RFMs capability to estimate anomalies is somewhat lower than capturing the seasonal cycle. For AnomalyR^2 the difference between the RFM and the LSMs is less pronounced.

To assess whether model performance is dependent on climate conditions, a correlation analysis was conducted, relating the spatial patterns in model performance to annual means of runoff, precipitation and temperature. Overall the results (Figure 8) indicate that there is little influence of mean climate on model performance (all correlations being $|r| < 0.5$). Nevertheless Figure 8 also suggests that there is some dependence of the relative bias on mean annual runoff. In addition Figure 8 suggests a possible link between mean temperature and Anomaly-R^2.

Finally, the difference between the cross validation in time and the cross validation in space is interesting to note. Overall the RFM has a slightly higher performance for the cross validation in space. This shows that the RFM is more skilful in estimating runoff dynamics at ungauged locations than at times without observations.

4.2.2 Basin scale validation

Although the RFM was initially developed to estimate grid-scale runoff it can also be used to derive first-order estimates of monthly river discharge. For this, monthly runoff from all grid cells within a river basin are spatially averaged for each time step. The resulting series of estimated monthly river discharge correspond reasonably well to the observed values (Figures 9 and 10). The RFM is also closer to the observations than the considered LSMs with respect to the majority of the performance metrics (S_{seas}, MEf, R^2 and Anomaly-R^2). However, in most river basins, two LSMs show as similar, ability in capturing the seasonal cycle of river discharge (Climatology-R^2) and the RFM is outperformed by the LSMs with respect to the relative bias.

4.2.3 Long-term mean evapotranspiration

The long-term difference between the WFD precipitation and RFM runoff was compared to a benchmark estimate of land evapotranspiration from the LandFlux-EVAL synthesis product (Mueller et al., 2013). Figure 11 shows the long-term mean evapotranspiration derived from the RFM and the LandFlux-EVAL synthesis product. Overall the two products agree well (R^2=0.66), and the RFM-based estimate lies in the majority of the cases within the uncertainty bounds of the LandFlux-EVAL product. Note that the RFM estimate does have small negative values in some parts of Scandinavia~~, which is related to a previously documented biasin the precipitation forcing (Gudmundsson et al., 2012b; Kauffeldt et al., 2013) .~~ . In these cases, the average amount of runoff predicted by the RFM is larger than the average precipitation provided by the WFD. This inconsistency does already emerge for the raw data entering the analysis. This has already been noted by Gudmundsson et al. (2012b) who found that the long-term mean runoff coefficient $\bar{Q}/\bar{P}$ computed from streamflow observations and WFD precipitation in Scandinavia are larger than one, suggesting that the WATCH forcing data underestimate precipitation in this region. In addition Kauffeldt et al. (2013) conclude that gridded precipitation products often underestimate precipitation in regions affected by snow undercatch. The RFM is able to account for this bias, and consequently the long-term mean difference between the WFD precipitation and estimated runoff is negative. This artefact is an intrinsic feature of the forcing data and only the development of unbiased precipitation estimates would resolve this inconsistency.

4.3 Example applications

4.3.1 Drought Monitoring

The RFM based gridded runoff estimates can for example be used to monitor surface water availability in Europe. While the monthly resolution may limit its ability to capture flash floods, it is still suitable for observing slowly evolving phenomena that are relevant for water resources management such as droughts. In Europe, 1976 is documented as a year with one of the most severe droughts of the twentieth century (Zaidman et al., 2002; Briffa et al., 2009; Tallaksen and Stahl, 2014). The severity of this drought is illustrated in Figure 12. Overall the runoff rates are low in large parts of Europe reaching values well below 1 mm day^{-1}. Accordingly monthly standardised runoff anomalies are negative in most parts of the continent and the extreme departures from normal conditions in southern England, France and central Europe corresponds to previously reported observations (Zaidman et al., 2002). As in Zaidman et al. (2002), runoff rates were log-transformed before standardisation, to account for the skewed distribution of the data.

4.3.2 A runoff climatology for Europe

Figure 13 shows a runoff climatology for Europe, that is based on the RFM based runoff estimates. The spatial pattern of the mean annual runoff rates highlights regions with abundant water availability in Central and Northern Europe. These are contrasted by low runoff rates in Southern and Eastern Europe. The maps displaying the month with the maximum and the month with the minimum of the mean annual cycle capture the contrasting influence of snow and evapotranspiration dynamics on runoff in Europe. On the one hand, snow accumulation leads to low flows in the winter months of the cold regions (high latitudes and high altitudes) and corresponding spring floods when the water stored as snow is released. On the other hand evapotranspiration rates follow the seasonality of the atmospheric water demand, leading to minimum runoff rates throughout late summer in large parts of central and southern Europe and winter floods in the West of the continent.

5 Discussion

5.1 Model selection and overfitting

The fact that increasing the model complexity, from a model that considers only atmospheric forcing (Equation (2)) to a model taking land parameters into account (Equation (3)), deteriorates model performance points towards issues with overfitting. Overfitting is referred to instances where the statistical model is fitted to random fluctuations (errors) instead of the true underlying relationship. This in turn leads to a reduction of the predictive power of the resulting model. As any machine learning technique, Random Forests are prone to overfitting, most likely in instances where the number of input variables that have no explanatory power increases (Hastie et al., 2009). In the context of this study, the fact that the inclusion of selected land parameters deteriorates the models performance therefore suggests that they have little or no explanatory power for continental scale runoff dynamics.

5.2 Model performance

The reasonable performance of the selected RFM with respect to (1) grid cell runoff, (2) discharge from continental drainage basins and (3) large-scale evapotranspiration demonstrates the fidelity of the RFM, also out of its expected comfort zone. The results from the cross validation show that the performance of the RFM reaches satisfactory levels, indicating that the employed technique is suitable for estimating monthly runoff at ungauged locations. Despite the fact that the selected RFM does not consider locally varying land parameters, the median performance measures lie within the range of other studies focusing on the prediction of monthly runoff at ungauged locations (Duan et al., 2006; Xia et al., 2012; Kumar et al., 2013; Blöschl et al., 2013).

The fact the RFM outperformed the considered LSMs with respect to most performance metrics (Figures 6, 7 and 10) shows that the RFM-based runoff estimates are closer to the observations than the considered LSMs with the exception of its mean bias. This possibly indicates that the considered LSMs have been optimised with respect to the mean continental river discharge, which might have introduced compensating errors in other features such as the seasonal cycle. Albeit a full explanation of the generally low performance of the LSMs lies beyond the scope of this study, it is also noteworthy that the differences between the RFM and the LSMs are most pronounced for the correlation between the observed and modelled mean seasonal cycles (R^2_{clim}). This issue has been previously reported (Gudmundsson et al., 2012b) and suggests that the LSMs may have deficiencies in capturing processes that govern the seasonality of runoff, such as evapotranspiration and snow dynamics.

5.3 Factors dominating large-scale terrestrial water dynamics

The results of the model selection procedure (Figure 4) do not only allow to identify the model setup that is best suited for estimating gridded monthly runoff in Europe, but also provide interesting clues on the optimal description of large-scale terrestrial water dynamics. The finding that the model forced by precipitation and temperature only is outperformed by the model considering the full atmospheric forcing, highlights the importance of the remaining atmospheric variables on terrestrial water dynamics. Among the factors that are likely to be important are the snowfall rate and drivers of evapotranspiration (e.g. radiation, humidity and wind speed). Nevertheless, the performance difference between these two modelling setups is relatively small if compared to the performance of the LSMs. This shows that gridded precipitation and temperature may be sufficient for estimating continental scale runoff dynamics with a reasonable degree of accuracy.

It is surprising that the inclusion of location specific land parameters did not improve the gridded runoff estimate. The fact that the spatial cross validation errors of the models with and without land parameters (equations (3) and (2) respectively) is not distinguishable implies that the influence of soil texture and topography on monthly runoff could not be detected. This, combined with previous results showing that signatures of runoff dynamics (Gudmundsson et al., 2011b; Sawicz et al., 2011; Ye et al., 2012; Yaeger et al., 2012; Szolgayova et al., 2014) as well as calibrated model parameters (van Werkhoven et al., 2008; Merz et al., 2011) are primarily controlled by climatic conditions, raises questions on the influence of location specific land parameters. In other words, one could speculate that the control of local variations of land parameters on large-scale terrestrial water dynamics may not be detectable, as their influence is overruled by atmospheric forcing. This is discussed in more detail in the following section.

5.4 Scale dependency and implications for model development

The fact that the influence of the considered land parameters did not improve the skill of the presented model raises interesting questions regarding the role of locally varying land parameters on terrestrial water dynamics. A likely explanation of this feature is related to the spatiotemporal resolution at which the machine learning model is applied, i.e. that locally varying land parameters may only have a minor influence on regional scale water fluxes. Previous publications have already suggested that the influence of land cover change on floods and droughts is more pronounced on small scales (e.g Blöschl et al., 2007) and that locally varying parameters do only have a minor influence on regional scale soil moisture simulations (e.g. Robock et al., 1998). Similarly Oudin et al. (2008) found only a weak empirical influence of land cover on longterm mean annual streamflow. However, an exhaustive assessment of such scale effects is still lacking.

While a complete assessment lies beyond the scope of this study a simple analysis of scale can provide some clues on the spatial and temporal resolution at which the effects of locally varying land parameters on runoff are expected to be detectable. For this we adopt the idea that terrestrial water dynamics has two separate space and time scales: A short scale where heterogeneous land properties dominate water dynamics and a large scale where homogeneous features of atmospheric forcing are dominating. Following previous suggestions (Vinnikov et al., 1996; Robock et al., 1998; Entin et al., 2000), the separation of time scales can be expressed as a mixture of two autocorrelation functions with exponential decay such that

$$r(\tau) = \zeta \exp\left(-\frac{\tau}{T_L}\right) + (1-\zeta)\exp\left(-\frac{\tau}{T_A}\right) \tag{8}$$

where τ is a time lag; the de-correlation time T_L is the time scale related to heterogeneous land properties, T_A the time scale related to the atmospheric forcing and $\zeta \,\epsilon\, [0,1]$ is the fraction of variance related to T_L. Note also that $T_L < T_A$. Similarly the separation of space scales can be expressed as

$$r(\lambda) = \eta \exp\left(-\frac{\lambda}{L_L}\right) + (1-\eta)\exp\left(-\frac{\lambda}{L_A}\right) \tag{9}$$

where λ is the lag distance, L_L is the length scale related to heterogeneous land properties, L_A the length scale related to the atmospheric forcing and $\eta \, \epsilon \, [0,1]$ is the fraction of variance related to L_L.

While the abovementioned separation of scales has been developed and is well documented for soil moisture (Vinnikov et al., 1996; Robock et al., 1998; Entin et al., 2000; Crow et al., 2012; Mittelbach and Seneviratne, 2012), its validity for other variables is less clear. Therefore we asses the applicability of Equations (8) and (9) for the considered streamflow observations in Europe. (Details on the estimation of space and time scales are summarised in Appendix A.) Figure 14 shows the estimated temporal and spatial correlation functions for runoff in Europe and Table 2 reports the parameters of Equations (8) and (9) fitted to the data. Overall, the small p-values of all parameters show that the hypothesised separation of scales is supported by observations. The time scale related to heterogeneous land parameters, T_L is approximately one week, which is well below the monthly resolution of the statistical model presented in this study. Similarly, the length scale related to land parameters L_L, is found to be $\leq$ 10 km, being substantially smaller than the edge length of the 0.5° grid cells. The results of this analysis of scales hence suggest that the effect of small scale variations in land parameters on runoff dynamics may only be detectable for models with spatial and temporal resolutions much higher than the one considered in this study. This is also consistent with the results of the model identification procedure, which could not find a significant improvement of model performance with to the inclusion of land parameters for the considered, coarse, spatiotemporal resolution.

6 Conclusions and Outlook

This study introduced a framework for estimating runoff on regular space-time grids in large spatial domains. The framework is based on the assumption that runoff at any location in space can be modelled as a function of gridded predictors, including both atmospheric variables and land parameters. While the framework has been applied to estimate monthly runoff on a 0.5° grid in Europe it can in principle be applied to finer spatial and temporal resolutions. The results from both model selection and model validation show that the model is capable to estimate monthly runoff dynamics at locations that were not used for model identification with a reasonable degree of accuracy. These results also show that the derived data are consistent with other variables of the terrestrial water cycle, which increases the confidence in the validity of the gridded runoff estimates. Such grids do allow to map historical runoff dynamics, providing first order estimates on its past evolution at any location in space, even if no ground observations are available. This is for example interesting in regions where no regular updates of streamflow archives exists (for Europe see e.g. Viglione et al., 2010). In such regions one could exploit the presented methodology to provide estimates of runoff for the years in which the station observations are not yet available.

Although the skill of the proposed method is reasonable and in line with previously published results (Duan et al., 2006; Xia et al., 2012; Kumar et al., 2013; Blöschl et al., 2013), there is still room for improving future estimates of runoff dynamics in Europe. Possible extensions of the presented analysis, each requiring an independent research effort, may focus on one of the following themes:

1. **Uncertainty of the considered data:** The considered atmospheric forcing data and the land parameters depend both on in-situ observations as well as on the methods used to derive estimates of the respective variables on a regular spatial grid. Unfortunately the uncertainty of the observations and the estimation procedures is often not documented in sufficient detail. However, several studies suggest that both the choice of atmospheric forcing data and mapped land parameters (e.g. Teuling et al., 2009; Guillod et al., 2013) can have pronounced impacts on simulation results. Similarly uncertainty estimates of the considered streamflow observations are not available.

2. **Limitations of the employed statistical methods:** Although Random Forests, like other machine learning techniques, are powerful tools for data driven modelling their application in the presented context may be limited. As other machine learning techniques they are prone to over fitting, implying that noise in the data can obscure possible signals (Hastie et al., 2009). Further, Random Forests do not explicitly handle spatial and temporal correlation in the data, and the implicit treatment of temporal correlations in equations (1) to (3) may be not sufficient. Consequently the application of other statistical techniques may improve large-scale estimates of terrestrial water dynamics in the future. Such work could potentially be based on top-kriging approaches (Sauquet et al., 2000; Skøien et al., 2006; Skøien and Blöschl, 2007; Laaha et al., 2013), that account for spatial dependence within the constraints of a channel network.

3. **Usage of large river basins for model identification:** This study did rely solely on streamflow from small catchments to estimate runoff at the grid-cell scale. However, discharge from large river basins does also carry information, which would be valuable to include into estimates of terrestrial water dynamics. A possible approach for this would be to first route the gridded runoff estimates through a channel network, and subsequently applying the procedure suggested by Fekete et al. (2002, 2011) to account for observations from large river basins.

4. **The non exhaustive list of considered land parameters:** In this study only the grid-cell slope and information on median grid-cell soil texture were taken into account. Although similar information is regularly used in LSMs, other parameters including the topographic index (Beven and Kirkby, 1979) or information on vegetation structure (Bonan, 2008) may have detectable impacts on large scale runoff dynamics in Europe.

5. **Temporal and spatial resolution:** The presented analysis is limited to relatively coarse spatial (0.5°) and temporal (monthly) resolution, focusing on large-scale phenomena. Obviously this

resolution limits the application of the derived data to the analysis of large, continental scale patterns. To which degree the suggested methodology is capable of capturing small scale variations of runoff (e.g. flash floods) remains an open question. Further investigations may help to clarify the effect of increasing the spatial and the temporal resolution on modelling runoff at ungauged locations using machine learning tools.

6. **Implications for model development:** The results from the model identification and validation raised interesting questions regarding the influence of land parameters on continental scale runoff dynamics. This, paired with an analysis of scales suggested that the influence of land parameters may only be detectable at model resolutions shorter than one week and smaller than ten kilometres. While this is consistent with the long history of catchment scale studies, it also raises questions on the optimal design of global scale models that are build to capture climatological phenomena. In fact, the results suggest that parsimonious physical descriptions, neglecting the influence of small scale variations in land parameters, may be sufficient to effectively describe terrestrial water dynamics on large scales. In a more formal setting, this can also be expressed as the hypothesis that hydrological variability at any location in space does solely depend on present and past atmospheric forcing – and not on locally varying land parameters. Of course this "Constant Land Parameter Hypothesis" (CLPH) will only be valid in certain circumstances and thus can act as a null hypothesis for testing the influence of selected land parameters on terrestrial water dynamics. This could guide the development efficient model physics.

In conclusion, we presented a novel approach for estimating the historical space-time evolution of runoff on regular spatial grids. The proposed methodology relies on the power of machine learning techniques to combine in-situ observations of runoff with gridded atmospheric variables. For Europe, the resulting runoff estimates compare well with observations and are consistent with other variables of the terrestrial water cycle, including evapotranspiration. Despite some remaining open questions, related e.g. to data uncertainty and spatiotemporal resolution, the derived runoff grid enables a new perspective on features of terrestrial water dynamics that emerge on large spatial scales. This was exemplified by (1) the validation of process based models, (2) the continuous mapping of runoff climatologies and (3) the analysis of hydrological droughts on large scales. Consequently, the resulting data product allows for a more comprehensive assessment of the historical space-time evolution of runoff in Europe relaxing the constraints of a limited observation network.

Appendix A

Estimating space and time scales of streamflow

Following a previous study (Skøien et al., 2003), daily streamflow observations from all catchments were log transformed and seasonal effects were removed. The deseasonalisation strictly follows recommendations on an removal of the seasonal cycle in the mean and the variance using harmonic regression (Hipel and McLeod, 1994; McLeod and Gweon, 2013). Temporal correlation was first estimated for each gauging station separately. The maximum time lag was limited to 120 days to reduce effects of climate induced interannual variability, which is reportedly strong in the data under investigation (Gudmundsson et al., 2011b). The estimated temporal autocorrelation functions from the individual stations were finally averaged as in previous studies (Entin et al., 2000; Skøien et al., 2003; Vinnikov et al., 1996) to obtain an estimate of the mean autocorrelation function of runoff in Europe. Spatial correlation was estimated using Morans I (Moran, 1950; Legendre and Legendre, 1998) for each time step separately with a spatial bin width of 10 km. This bin width is a compromise between having enough station pairs per bin and the ability to resolve small scale processes (the first bin contains 31 pairs, the median number of pairs: 490). The analysis of spatial correlation was limited to a maximum lag distance of 400 kilometres to reduce the effect of large scale climate gradients, which impact European runoff dynamics (Gudmundsson et al., 2011a,b). Finally the spatial correlation functions were then averaged over all time steps, resulting in an estimate of mean spatial correlation for the time period under investigation.

Acknowledgements. This research contributes to the European Union (FP7) funded project DROUGHT-R&SPI (contract no. 282769). Partial support of the ERC DROUGHT-HEAT project (contract no. 617518) is acknowledged. The effort to assemble the European Water Archive (EWA) by of the UNESCOIHP VII FRIEND programme, the data management by the GRDC, the generation of the WFD and the LSM ensemble by members of the European Union WATCH project (FP6) are gratefully acknowledged.

References

Arnell, N. W.: Grid mapping of river discharge, Journal of Hydrology, 167, 39 – 56, doi:10.1016/0022-1694(94)02626-M, 1995.

Baldocchi, D.: TURNER REVIEW No. 15. ”Breathing” of the terrestrial biosphere: lessons learned from a global network of carbon dioxide flux measurement systems, Aust. J. Bot., 56, 1–26, doi:10.1071/BT07151, 2008.

Balsamo, G., Albergel, C., Beljaars, A., Boussetta, S., Cloke, H., Dee, D., Dutra, E., Muñoz Sabater, J., Pappenberger, F., de Rosnay, P., Stockdale, T., and Vitart, F.: ERA-Interim/Land: a global land water resources dataset, Hydrology and Earth System Sciences Discussions, 10, 14 705–14 745, doi:10.5194/hessd-10-14705-2013, 2013.

Beck, H. E., van Dijk, A. I. J. M., Miralles, D. G., de Jeu, R. A. M., (Sampurno) Bruijnzeel, L. A., McVicar, T. R., and Schellekens, J.: Global patterns in base flow index and recession based on streamflow observations from 3394 catchments, Water Resources Research, 49, 7843–7863, doi:10.1002/2013WR013918, 2013.

Beven, K. J. and Kirkby, M. J.: A physically based, variable contributing area model of basin hydrology, Hydrological sciences journal, 24, 43–69, doi:10.1080/02626667909491834, 1979.

Bishop, C. M.: Pattern Recognition and Machine Learning, Information Science and Statistics, Springer, 2006.

Blöschl, G., Ardoin-Bardin, S., Bonell, M., Dorninger, M., Goodrich, D., Gutknecht, D., Matamoros, D., Merz, B., Shand, P., and Szolgay, J.: At what scales do climate variability and land cover change impact on flooding and low flows?, Hydrological Processes, 21, 1241–1247, doi:10.1002/hyp.6669, 2007.

Blöschl, G., Sivapalan, M., Wagener, T., Viglione, A., and Savenije, H., eds.: Runoff Prediction in Ungauged Basins: Synthesis Across Processes, Places and Scales, Cambridge University Press, 2013.

Bonan, G. B.: Ecological Climatology: Concepts and Applications, Cambridge University Press, 2 edn., 2008.

Breiman, L.: Random Forests, Machine Learning, 45, 5–32, doi:10.1023/A:1010933404324, 2001.

Briffa, K. R., van der Schrier, G., and Jones, P. D.: Wet and dry summers in Europe since 1750: evidence of increasing drought, International Journal of Climatology, 29, 1894–1905, doi:10.1002/joc.1836, 2009.

Ciais, P., Reichstein, M., Viovy, N., Granier, A., Ogee, J., Allard, V., Aubinet, M., Buchmann, N., Bernhofer, C., Carrara, A., Chevallier, F., De Noblet, N., Friend, A. D., Friedlingstein, P., Grunwald, T., Heinesch, B., Keronen, P., Knohl, A., Krinner, G., Loustau, D., Manca, G., Matteucci, G., Miglietta, F., Ourcival, J. M., Papale, D., Pilegaard, K., Rambal, S., Seufert, G., Soussana, J. F., Sanz, M. J., Schulze, E. D., Vesala, T., and Valentini, R.: Europe-wide reduction in primary productivity caused by the heat and drought in 2003, Nature, 437, 529–533, doi:10.1038/nature03972, 2005.

Crow, W. T., Berg, A. A., Cosh, M. H., Loew, A., Mohanty, B. P., Panciera, R., de Rosnay, P., Ryu, D., and Walker, J. P.: Upscaling sparse ground-based soil moisture observations for the validation of coarse-resolution satellite soil moisture products, Reviews of Geophysics, 50, RG2002, doi:10.1029/2011RG000372, 2012.

de Jeu, R., Wagner, W., Holmes, T., Dolman, A., van de Giesen, N., and Friesen, J.: Global Soil Moisture Patterns Observed by Space Borne Microwave Radiometers and Scatterometers, Surveys in Geophysics, 29, 399–420, doi:10.1007/s10712-008-9044-0, 10.1007/s10712-008-9044-0, 2008.

de Linage, C., Famiglietti, J. S., and Randerson, J. T.: Statistical prediction of terrestrial water storage changes in the Amazon Basin using tropical Pacific and North Atlantic sea surface temperature anomalies, Hydrology and Earth System Sciences, 18, 2089–2102, doi:10.5194/hess-18-2089-2014, 2014.

Dettinger, M. and Diaz, H.: Global Characteristics of Stream Flow Seasonality and Variability, Journal of Hydrometeorology, 1, 289–310, doi:10.1175/1525-7541(2000)001<0289:GCOSFS>2.0.CO;2, 2000.

Dirmeyer, P. A., Gao, X., Zhao, M., Guo, Z., Oki, T., and Hanasaki, N.: GSWP-2: Multimodel Analysis and Implications for Our Perception of the Land Surface, Bulletin of the American Meteorological Society, 87, 1381–1397, doi:10.1175/BAMS-87-10-1381, 2006.

Döll, P., Fiedler, K., and Zhang, J.: Global-scale analysis of river flow alterations due to water withdrawals and reservoirs, Hydrology and Earth System Sciences, 13, 2413–2432, doi:10.5194/hess-13-2413-2009, 2009.

Dorigo, W., Xaver, A., Vreugdenhil, M., Gruber, A., Hegyiová, A., Sanchis-Dufau, A., Zamojski, D., Cordes, C., Wagner, W., and Drusch, M.: Global Automated Quality Control of In Situ Soil Moisture Data from the International Soil Moisture Network, Vadose Zone Journal, 12, 21, doi:10.2136/vzj2012.0097, 2013.

Duan, Q., Schaake, J., Andréassian, V., Franks, S., Goteti, G., Gupta, H., Gusev, Y., Habets, F., Hall, A., Hay, L., Hogue, T., Huang, M., Leavesley, G., Liang, X., Nasonova, O., Noilhan, J., Oudin, L., Sorooshian, S., Wagener, T., and Wood, E.: Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops, Journal of Hydrology, 320, 3 – 17, doi:10.1016/j.jhydrol.2005.07.031, the model parameter estimation experiment - MOPEX, 2006.

Entin, J. K., Robock, A., Vinnikov, K. Y., Hollinger, S. E., Liu, S., and Namkhai, A.: Temporal and spatial scales of observed soil moisture variations in the extratropics, J. Geophys. Res., 105, 865 – 877, doi:10.1029/2000JD900051, 2000.

Famiglietti, J. S. and Rodell, M.: Water in the Balance, Science, 340, 1300–1301, doi:10.1126/science.1236460, 2013.

FAO, IIASA, ISRIC, ISSCAS, and JRC: Harmonized World Soil Database (version 1.2), Tech. rep., FAO, Rome, Italy and IIASA, Laxenburg, Austria, 2012.

Fekete, B., Maurer, T., and Vörösmarty, C. J.: ISLSCP II UNH/GRDC Composite Monthly Runoff, in: ISLSCP Initiative II Collection. Data set., edited by Hall, I., G., F., Collatz, G., Meeson, B., Los, S., de Colstoun, E. B., and Landis, D., Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A, doi:10.3334/ORNLDAAC/994, 2011.

Fekete, B. M., Vörösmarty, C. J., and Grabs, W.: High-resolution fields of global runoff combining observed river discharge and simulated water balances, Global Biogeochem. Cycles, 16, 1042, doi:10.1029/1999GB001254, 2002.

Fekete, B. M., Looser, U., Pietroniro, A., and Robarts, R. D.: Rationale for Monitoring Discharge on the Ground, J. Hydrometeor, 13, 1977–1986, doi:10.1175/JHM-D-11-0126.1, 2012.

Granier, A., Reichstein, M., Bréda, N., Janssens, I., Falge, E., Ciais, P., Grünwald, T., Aubinet, M., Berbigier, P., Bernhofer, C., Buchmann, N., Facini, O., Grassi, G., Heinesch, B., Ilvesniemi, H., Keronen, P., Knohl, A., Köstner, B., Lagergren, F., Lindroth, A., Longdoz, B., Loustau, D., Mateus, J., Montagnani, L., Nys, C., Moors, E., Papale, D., Peiffer, M., Pilegaard, K., Pita, G., Pumpanen, J., Rambal, S., Rebmann, C., Rodrigues, A., Seufert, G., Tenhunen, J., Vesala, T., and Wang, Q.: Evidence for soil water control on carbon and water dynamics in European forests during the extremely dry year: 2003, Agricultural and Forest Meteorology, 143, 123 – 145, doi:10.1016/j.agrformet.2006.12.004, 2007.

Greve, P., Orlowsky, B., Mueller, B., Sheffield, J., Reichstein, M., and Seneviratne, S. I.: Global assessment of trends in wetting and drying over land, Nature Geosci, advance online publication, doi:10.1038/ngeo2247, 2014.

Guan, K., Pan, M., Li, H., Wolf, A., Wu, J., Medvigy, D., Caylor, K. K., Sheffield, J., Wood, E. F., Malhi, Y., Liang, M., Kimball, J. S., Saleska, S. R., Berry, J., Joiner, J., and Lyapustin, A. I.: Photosynthetic seasonality of global tropical forests constrained by hydroclimate, Nature Geosci, advance online publication, –, doi:10.1038/ngeo2382, 2015.

Gudmundsson, L. and Seneviratne, S. I.: A comprehensive drought climatology for Europe (1950 –2013), in: Drought: Research and Science-Policy Interfacing, pp. 31–37, CRC Press, doi:10.1201/b18077-7, 2015.

Gudmundsson, L., Tallaksen, L. M., and Stahl, K.: Spatial cross-correlation patterns of European low, mean and high flows, Hydrological Processes, 25, 1034–1045, doi:10.1002/hyp.7807, 2011a.

Gudmundsson, L., Tallaksen, L. M., Stahl, K., and Fleig, A. K.: Low-frequency variability of European runoff, Hydrology and Earth System Sciences, 15, 2853–2869, doi:10.5194/hess-15-2853-2011, 2011b.

Gudmundsson, L., Tallaksen, L. M., Stahl, K., Clark, D. B., Dumont, E., Hagemann, S., Bertrand, N., Gerten, D., Heinke, J., Hanasaki, N., Voss, F., and Koirala, S.: Comparing Large-Scale Hydrological Model Simulations to Observed Runoff Percentiles in Europe, J. Hydrometeor, 13, 604–620, doi:10.1175/JHM-D-11-083.1, 2012a.

Gudmundsson, L., Wagener, T., Tallaksen, L. M., and Engeland, K.: Evaluation of nine large-scale hydrological models with respect to the seasonal runoff climatology in Europe, Water Resour. Res., 48, W11 504, doi:10.1029/2011WR010911, 2012b.

Gudmundsson, L., Rego, F. C., Rocha, M., and Seneviratne, S. I.: Predicting above normal wildfire activity in southern Europe as a function of meteorological drought, Environmental Research Letters, 9, 084 008, doi:10.1088/1748-9326/9/8/084008, 2014.

Guillod, B., Davin, E., Kündig, C., Smiatek, G., and Seneviratne, S. I.: Impact of soil map specifications for European climate simulations, Climate Dynamics, 40, 123–141, doi:10.1007/s00382-012-1395-z, 2013.

Haddeland, I., Clark, D. B., Franssen, W., Ludwig, F., Voß, F., Arnell, N. W., Bertrand, N., Best, M., Folwell, S., Gerten, D., Gomes, S., Gosling, S. N., Hagemann, S., Hanasaki, N., Harding, R., Heinke, J., Kabat, P., Koirala, S., Oki, T., Polcher, J., Stacke, T., Viterbo, P., Weedon, G. P., and Yeh, P.: Multimodel Estimate of the Global Terrestrial Water Balance: Setup and First Results, Journal of Hydrometeorology, 12, 869–884, doi:10.1175/2011JHM1324.1, 2011.

Haddeland, I., Heinke, J., Voß, F., Eisner, S., Chen, C., Hagemann, S., and Ludwig, F.: Effects of climate model radiation, humidity and wind estimates on hydrological simulations, Hydrology and Earth System Sciences, 16, 305–318, doi:10.5194/hess-16-305-2012, 2012.

Hannah, D. M., Demuth, S., van Lanen, H. A. J., Looser, U., Prudhomme, C., Rees, G., Stahl, K., and Tallaksen, L. M.: Large-scale river flow archives: importance, current status and future needs, Hydrological Processes, 25, 1191–1200, doi:10.1002/hyp.7794, 2011.

Hastie, T., Tibshirani, R., and Friedman, J. H.: The Elements of Statistical Learning – Data Mining, Inference, and Prediction, Second Edition, Springer Series in Statistics, Springer, 2 edn., 2009.

Hipel, K. and McLeod, A.: Time series modelling of water resources and environmental systems, vol. 45 of *Developments in Water Science*, Elsevier, 1994.

Hirschi, M., Seneviratne, S. I., and Schär, C.: Seasonal Variations in Terrestrial Water Storage for Major Midlatitude River Basins, J. Hydrometeor, 7, 39–60, doi:10.1175/JHM480.1, 2006.

Hirschi, M., Mueller, B., and Seneviratne, S. I.: Using remotely sensed soil moisture for land-atmosphere coupling diagnostics: The role of surface vs. root-zone soil moisture variability, Remote Sensing of Environment, 154, 246 – 252, doi:10.1016/j.rse.2014.08.030, 2014.

Houborg, R., Rodell, M., Li, B., Reichle, R., and Zaitchik, B. F.: Drought indicators based on model-assimilated Gravity Recovery and Climate Experiment (GRACE) terrestrial water storage observations, Water Resour. Res., 48, W07 525, doi:10.1029/2011WR011291, 2012.

Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., Arheimer, B., Blume, T., Clark, M. P., Ehret, U., Fenicia, F., Freer, J. E., Gelfan, A., Gupta, H. V., Hughes, D. A., Hut, R. W., Montanari, A., Pande, S., Tetzlaff, D., Troch, P. A., Uhlenbrook, S., Wagener, T., Winsemius, H. C., Woods, R. A., Zehe, E., and Cudennec, C.: A decade of Predictions in Ungauged Basins (PUB) – a review, Hydrological Sciences Journal, 58:6, 1198 – 1255, doi:10.1080/02626667.2013.803183, 2013.

Jung, M., Reichstein, M., and Bondeau, A.: Towards global empirical upscaling of FLUXNET eddy covariance observations: validation of a model tree ensemble approach using a biosphere model, Biogeosciences, 6, 2001–2013, doi:10.5194/bg-6-2001-2009, 2009.

Jung, M., Reichstein, M., Ciais, P., Seneviratne, S. I., Sheffield, J., Goulden, M. L., Bonan, G., Cescatti, A., Chen, J., de Jeu, R., Dolman, A. J., Eugster, W., Gerten, D., Gianelle, D., Gobron, N., Heinke, J., Kimball, J., Law, B. E., Montagnani, L., Mu, Q., Mueller, B., Oleson, K., Papale, D., Richardson, A. D., Roupsard, O., Running, S., Tomelleri, E., Viovy, N., Weber, U., Williams, C., Wood, E., Zaehle, S., and Zhang, K.: Recent decline in the global land evapotranspiration trend due to limited moisture supply, Nature, 467, 951 – 954, doi:10.1038/nature09396, 2010.

Jung, M., Reichstein, M., Margolis, H. A., Cescatti, A., Richardson, A. D., Arain, M. A., Arneth, A., Bernhofer, C., Bonal, D., Chen, J., Gianelle, D., Gobron, N., Kiely, G., Kutsch, W., Lasslop, G., Law, B. E., Lindroth, A., Merbold, L., Montagnani, L., Moors, E. J., Papale, D., Sottocornola, M., Vaccari, F., and Williams, C.: Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations, J. Geophys. Res., 116, G00J07, doi:10.1029/2010JG001566, 2011.

Kauffeldt, A., Halldin, S., Rodhe, A., Xu, C.-Y., and Westerberg, I. K.: Disinformative data in large-scale hydrological modelling, Hydrology and Earth System Sciences, 17, 2845–2857, doi:10.5194/hess-17-2845-2013, 2013.

Koster, R. D., Dirmeyer, P. A., Guo, Z., Bonan, G., Chan, E., Cox, P., Gordon, C. T., Kanae, S., Kowalczyk, E., Lawrence, D., Liu, P., Lu, C.-H., Malyshev, S., McAvaney, B., Mitchell, K., Mocko, D., Oki, T., Oleson, K., Pitman, A., Sud, Y. C., Taylor, C. M., Verseghy, D., Vasic, R., Xue, Y., and Yamada, T.: Regions of Strong Coupling Between Soil Moisture and Precipitation, Science, 305, 1138–1140, doi:10.1126/science.1100217, 2004.

Krakauer, N. Y. and Fung, I.: Mapping and attribution of change in streamflow in the coterminous United States, Hydrology and Earth System Sciences, 12, 1111–1120, doi:10.5194/hess-12-1111-2008, 2008.

Kumar, R., Livneh, B., and Samaniego, L.: Toward computationally efficient large-scale hydrologic predictions with a multiscale regionalization scheme, Water Resources Research, 49, 5700–5714, doi:10.1002/wrcr.20431, 2013.

Laaha, G., Skøien, J. O., Nobilis, F., and Bløschl, G.: Spatial Prediction of Stream Temperatures Using Top-Kriging with an External Drift, Environmental Modeling & Assessment, 18, 671–683, doi:10.1007/s10666-013-9373-3, 2013.

Landerer, F. W. and Swenson, S. C.: Accuracy of scaled GRACE terrestrial water storage estimates, Water Resour. Res., 48, W04 531, doi:10.1029/2011WR011453, 2012.

Legendre, P. and Legendre, L.: Numerical ecology, Elsevier New York, 1998.

Liaw, A. and Wiener, M.: Classification and Regression by randomForest, R News, 2, 18 – 22, 2002.

Loew, A., Stacke, T., Dorigo, W., de Jeu, R., and Hagemann, S.: Potential and limitations of multidecadal satellite soil moisture observations for selected climate model evaluation studies, Hydrology and Earth System Sciences, 17, 3523–3542, doi:10.5194/hess-17-3523-2013, 2013.

McLeod, A. I. and Gweon, H.: Optimal Deseasonalization for Monthly and Daily Geophysical Time Series, Journal of Environmentl Statistics, 4, http://jes.stat.ucla.edu/v04/i11, 2013.

Merz, R., Parajka, J., and Blöschl, G.: Time stability of catchment model parameters: Implications for climate impact analyses, Water Resour. Res., 47, W02 531, doi:10.1029/2010WR009505, 2011.

Miralles, D. G., van den Berg, M. J., Gash, J. H., Parinussa, R. M., de Jeu, R. A. M., Beck, H. E., Holmes, T. R. H., Jiménez, C., Verhoest, N. E. C., Dorigo, W. A., Teuling, A. J., and Johannes Dolman, A.: El Niño-La Niña cycle and recent trends in continental evaporation, Nature Clim. Change, 4, 122–126, doi:10.1038/nclimate2068, 2014.

Mittelbach, H. and Seneviratne, S. I.: A new perspective on the spatio-temporal variability of soil moisture: temporal dynamics versus time-invariant contributions, Hydrology and Earth System Sciences, 16, 2169–2179, doi:10.5194/hess-16-2169-2012, 2012.

Moran, P. A. P.: Notes on Continuous Stochastic Phenomena, Biometrika, 37, 17 – 23, 1950.

Mueller, B. and Seneviratne, S. I.: Hot days induced by precipitation deficits at the global scale, Proceedings of the National Academy of Sciences, 109, 12 398 – 12 403, doi:10.1073/pnas.1204330109, 2012.

Mueller, B., Hirschi, M., and Seneviratne, S. I.: New diagnostic estimates of variations in terrestrial water storage based on ERA-Interim data, Hydrological Processes, 25, 996–1008, doi:10.1002/hyp.7652, 2011a.

Mueller, B., Seneviratne, S. I., Jimenez, C., Corti, T., Hirschi, M., Balsamo, G., Ciais, P., Dirmeyer, P., Fisher, J. B., Guo, Z., Jung, M., Maignan, F., McCabe, M. F., Reichle, R., Reichstein, M., Rodell, M., Sheffield, J., Teuling, A. J., Wang, K., Wood, E. F., and Zhang, Y.: Evaluation of global observations-based evapotranspiration datasets and IPCC AR4 simulations, Geophys. Res. Lett., 38, L06 402, doi:10.1029/2010GL046230, 2011b.

Mueller, B., Hirschi, M., Jimenez, C., Ciais, P., Dirmeyer, P. A., Dolman, A. J., Fisher, J. B., Jung, M., Ludwig, F., Maignan, F., Miralles, D. G., McCabe, M. F., Reichstein, M., Sheffield, J., Wang, K., Wood, E. F., Zhang, Y., and Seneviratne, S. I.: Benchmark products for land evapotranspiration: LandFlux-EVAL multi-data set synthesis, Hydrology and Earth System Sciences, 17, 3707–3720, doi:10.5194/hess-17-3707-2013, 2013.

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

Orlowsky, B., Hoekstra, A. Y., Gudmundsson, L., and Seneviratne, S. I.: Today's virtual water consumption and trade under future water scarcity, Environmental Research Letters, 9, 074 007, doi:10.1088/1748-9326/9/7/074007, 2014.

Oudin, L., Andréassian, V., Lerat, J., and Michel, C.: Has land cover a significant impact on mean annual streamflow? An international assessment using 1508 catchments, Journal of Hydrology, 357, 303 – 316, doi:10.1016/j.jhydrol.2008.05.021, 2008.

Parajka, J., Viglione, A., Rogger, M., Salinas, J. L., Sivapalan, M., and Blöschl, G.: Comparative assessment of predictions in ungauged basins &ndash; Part 1: Runoff-hydrograph studies, Hydrology and Earth System Sciences, 17, 1783–1795, doi:10.5194/hess-17-1783-2013, 2013.

Prudhomme, C., Giuntoli, I., Robinson, E. L., Clark, D. B., Arnell, N. W., Dankers, R., Fekete, B. M., Franssen, W., Gerten, D., Gosling, S. N., Hagemann, S., Hannah, D. M., Kim, H., Masaki, Y., Satoh, Y., Stacke, T., Wada, Y., and Wisser, D.: Hydrological droughts in the 21st century, hotspots and uncertainties from a global multimodel ensemble experiment, Proceedings of the National Academy of Sciences, 111, 3262–3267, doi:10.1073/pnas.1222473110, 2014.

Reager, J. and Famiglietti, J. S.: Characteristic mega-basin water storage behavior using GRACE, Water Resources Research, 49, 3314 – 3329, doi:10.1002/wrcr.20264, 2013.

Reichstein, M., Bahn, M., Ciais, P., Frank, D., Mahecha, M. D., Seneviratne, S. I., Zscheischler, J., Beer, C., Buchmann, N., Frank, D. C., Papale, D., Rammig, A., Smith, P., Thonicke, K., van der Velde, M., Vicca, S., Walz, A., and Wattenbach, M.: Climate extremes and the carbon cycle, Nature, 500, 287–295, doi:10.1038/nature12350, 2013.

Robock, A., Schlosser, C. A., Vinnikov, K. Y., Speranskaya, N. A., Entin, J. K., and Qiu, S.: Evaluation of the AMIP soil moisture simulations, Global and Planetary Change, 19, 181 – 208, doi:10.1016/S0921-8181(98)00047-2, 1998.

Rodell, M. and Famiglietti, J. S.: Detectability of variations in continental water storage from satellite observations of the time dependent gravity field, Water Resour. Res., 35, 2705–2723, doi:10.1029/1999WR900141, 1999.

Rodell, M., Houser, P. R., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C.-J., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., Entin*, J. K., Walker, J. P., Lohmann, D., and Toll, D.: The Global Land Data Assimilation System, Bull. Amer. Meteor. Soc., 85, 381–394, doi:10.1175/BAMS-85-3-381, 2004.

Sauquet, E., Krasovskaia, I., and Leblois, E.: Mapping mean monthly runoff pattern using EOF analysis, Hydrology and Earth System Sciences, 4, 79–93, doi:10.5194/hess-4-79-2000, 2000.

Sawicz, K., Wagener, T., Sivapalan, M., Troch, P. A., and Carrillo, G.: Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA, Hydrology and Earth System Sciences, 15, 2895–2911, doi:10.5194/hess-15-2895-2011, 2011.

Schewe, J., Heinke, J., Gerten, D., Haddeland, I., Arnell, N. W., Clark, D. B., Dankers, R., Eisner, S., Fekete, B. M., Colón-González, F. J., Gosling, S. N., Kim, H., Liu, X., Masaki, Y., Portmann, F. T., Satoh, Y., Stacke, T., Tang, Q., Wada, Y., Wisser, D., Albrecht, T., Frieler, K., Piontek, F., Warszawski, L., and Kabat, P.: Multimodel assessment of water scarcity under climate change, Proceedings of the National Academy of Sciences, 111, 3245–3250, doi:10.1073/pnas.1222460110, 2014.

Seneviratne, S. I., Corti, T., Davin, E. L., Hirschi, M., Jaeger, E. B., Lehner, I., Orlowsky, B., and Teuling, A. J.: Investigating soil moisture-climate interactions in a changing climate: A review, Earth-Science Reviews, 99, 125 – 161, doi:10.1016/j.earscirev.2010.02.004, 2010.

Sheffield, J., Wood, E. F., and Roderick, M. L.: Little change in global drought over the past 60 years, Nature, 491, 435–438, doi:10.1038/nature11575, 2012.

Sivapalan, M., Takeuchi, K., Franks, S. W., Gupta, V. K., Karambiri, H., Lakshmi, V., Liang, X., Mc Donnell, J. J., Mendiondo, E. M., O'Connell, P. E., Oki, T., Pomeroy, J. W., Schertzer, D., Uhlenbrook, S., and Zehe, E.: IAHS Decade on Predictions in Ungauged Basins (PUB), 2003 - 2012: Shaping an exciting future for the hydrological sciences, Hydrological Sciences Journal, 48, 857–880, doi:10.1623/hysj.48.6.857.51421, 2003.

Skøien, J. O. and Blöschl, G.: Spatiotemporal topological kriging of runoff time series, Water Resources Research, 43, W09 419, doi:10.1029/2006WR005760, 2007.

Skøien, J. O., Blöschl, G., and Western, A. W.: Characteristic space scales and timescales in hydrology, Water Resour. Res., 39, 1304, doi:10.1029/2002WR001736, 2003.

Skøien, J. O., Merz, R., and Blöschl, G.: Top-kriging - geostatistics on stream networks, Hydrology and Earth System Sciences, 10, 277–287, doi:10.5194/hess-10-277-2006, 2006.

Slack, J. R. and Landwehr, J. M.: Hydroclimatic Data Network (HCDN): A USGS Streamflow Data Set for the United States for the Study of Climate Variations, 1874-1988, USGS Open-File Report, 92, 129–193, http://pubs.usgs.gov/of/1992/ofr92-129/, 1992.

Stahl, K., Hisdal, H., Hannaford, J., Tallaksen, L. M., van Lanen, H. A. J., Sauquet, E., Demuth, S., Fendekova, M., and Jódar, J.: Streamflow trends in Europe: evidence from a dataset of near-natural catchments, Hydrology and Earth System Sciences, 14, 2367–2382, doi:10.5194/hess-14-2367-2010, 2010.

Stahl, K., Tallaksen, L. M., Hannaford, J., and van Lanen, H. A. J.: Filling the white space on maps of European runoff trends: estimates from a multi-model ensemble, Hydrology and Earth System Sciences, 16, 2035–2047, doi:10.5194/hess-16-2035-2012, 2012.

Szolgayova, E., Laaha, G., Blöschl, G., and Bucher, C.: Factors influencing long range dependence in streamflow of European rivers, Hydrological Processes, 28, 1573–1586, doi:10.1002/hyp.9694, 2014.

Tallaksen, L. M. and Stahl, K.: Spatial and temporal patterns of large-scale droughts in Europe: Model dispersion and performance, Geophysical Research Letters, 41, 429–434, doi:10.1002/2013GL058573, 2014.

Teuling, A. J., Uijlenhoet, R., van den Hurk, B., and Seneviratne, S. I.: Parameter Sensitivity in LSMs: An Analysis Using Stochastic Soil Moisture Models and ELDAS Soil Parameters, Journal of Hydrometeorology, 10, 751–765, doi:10.1175/2008JHM1033.1, 2009.

Thomas, A. C., Reager, J. T., Famiglietti, J. S., and Rodell, M.: A GRACE-based water storage deficit approach for hydrological drought characterization, Geophysical Research Letters, 41, 1537–1545, doi:10.1002/2014GL059323, 2014.

Tootle, G. A. and Piechota, T. C.: Relationships between Pacific and Atlantic ocean sea surface temperatures and U.S. streamflow variability, Water Resour. Res., 42, W07 411, doi:10.1029/2005WR004184, 2006.

van Werkhoven, K., Wagener, T., Reed, P., and Tang, Y.: Characterization of watershed model behavior across a hydroclimatic gradient, Water Resour. Res., 44, W01 429, doi:10.1029/2007WR006271, 2008.

Viglione, A., Borga, M., Balabanis, P., and Blöschl, G.: Barriers to the exchange of hydrometeorological data in Europe: Results from a survey and implications for data policy, Journal of Hydrology, 394, 63 – 77, doi:10.1016/j.jhydrol.2010.03.023, flash Floods: Observations and Analysis of Hydrometeorological Controls, 2010.

Vinnikov, K. Y., Robock, A., Speranskaya, N. A., and Schlosser, C. A.: Scales of temporal and spatial variability of midlatitude soil moisture, Journal of Geophysical Research, 101(D3), 7163 – 7174, doi:10.1029/95JD02753, 1996.

Vörösmarty, C. J., McIntyre, P. B., Gessner, M. O., Dudgeon, D., Prusevich, A., Green, P., Glidden, S., Bunn, S. E., Sullivan, C. A., Liermann, C. R., and Davies, P. M.: Global threats to human water security and river biodiversity, Nature, 467, 555–561, doi:10.1038/nature09440, 2010.

Wagner, W., Blöschl, G., Pampaloni, P., Calvet, J.-C., Bizzarri, B., Wigneron, J.-P., and Kerr, Y.: Operational readiness of microwave remote sensing of soil moisture for hydrologic applications, Nordic Hydrology, 38, 1 – 20, doi:10.2166/nh.2007.029, 2007.

Weedon, G. P., Gomes, S., Viterbo, P., Shuttleworth, W. J., Blyth, E., Österle, H., Adam, J. C., Bellouin, N., Boucher, O., and Best, M.: Creation of the WATCH Forcing Data and Its Use to Assess Global and Regional Reference Crop Evaporation over Land during the Twentieth Century, Journal of Hydrometeorology, 12, 823–848, doi:10.1175/2011JHM1369.1, 2011.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, vol. 100 of *International Geophysics Series*, Academic Press, 3 edn., 2011.

Wong, T. H. F. and Laurenson, E. M.: Wave speed-discharge relations in natural channels, Water Resources Research, 19, 701–706, doi:10.1029/WR019i003p00701, 1983.

Xia, Y., Mitchell, K., Ek, M., Cosgrove, B., Sheffield, J., Luo, L., Alonge, C., Wei, H., Meng, J., Livneh, B., Duan, Q., and Lohmann, D.: Continental-scale water and energy flux analysis and validation for North American Land Data Assimilation System project phase 2 (NLDAS-2): 2. Validation of model-simulated streamflow, Journal of Geophysical Research: Atmospheres, 117, D03 110, doi:10.1029/2011JD016051, 2012.

Yaeger, M., Coopersmith, E., Ye, S., Cheng, L., Viglione, A., and Sivapalan, M.: Exploring the physical controls of regional patterns of flow duration curves &ndash; Part 4: A synthesis of empirical analysis, process modeling and catchment classification, Hydrology and Earth System Sciences, 16, 4483–4498, doi:10.5194/hess-16-4483-2012, 2012.

Ye, S., Yaeger, M., Coopersmith, E., Cheng, L., and Sivapalan, M.: Exploring the physical controls of regional patterns of flow duration curves &ndash; Part 2: Role of seasonality, the regime curve, and associated process controls, Hydrology and Earth System Sciences, 16, 4447–4465, doi:10.5194/hess-16-4447-2012, 2012.

Zaidman, M. D., Rees, H. G., and Young, A. R.: Spatio-temporal development of streamflow droughts in north-west Europe, Hydrology and Earth System Sciences, 6, 733–751, doi:10.5194/hess-6-733-2002, 2002.

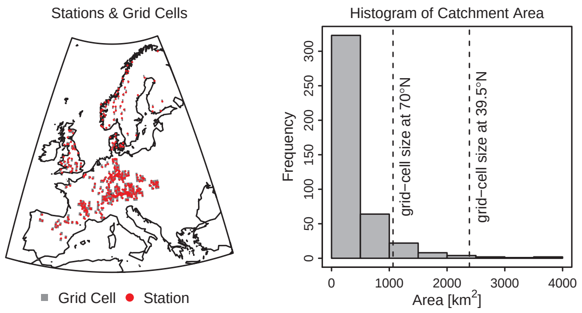

Fig. 1. Runoff observations: Left: Locations of the gauging stations of the considered catchments, as well as the grid cells with observations. Right: Histogram of catchment areas. The vertical lines indicate the grid-cell size of the southern- and northernmost grid cells.

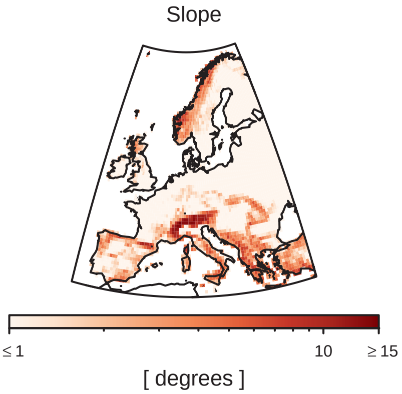

Fig. 2. Median grid-cell slope.

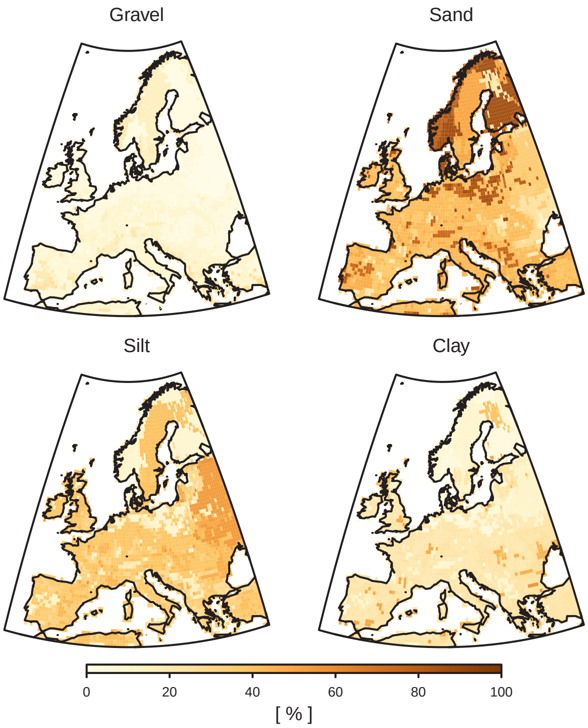

Fig. 3. Soil texture: median fraction of gravel, sand, silt and clay.

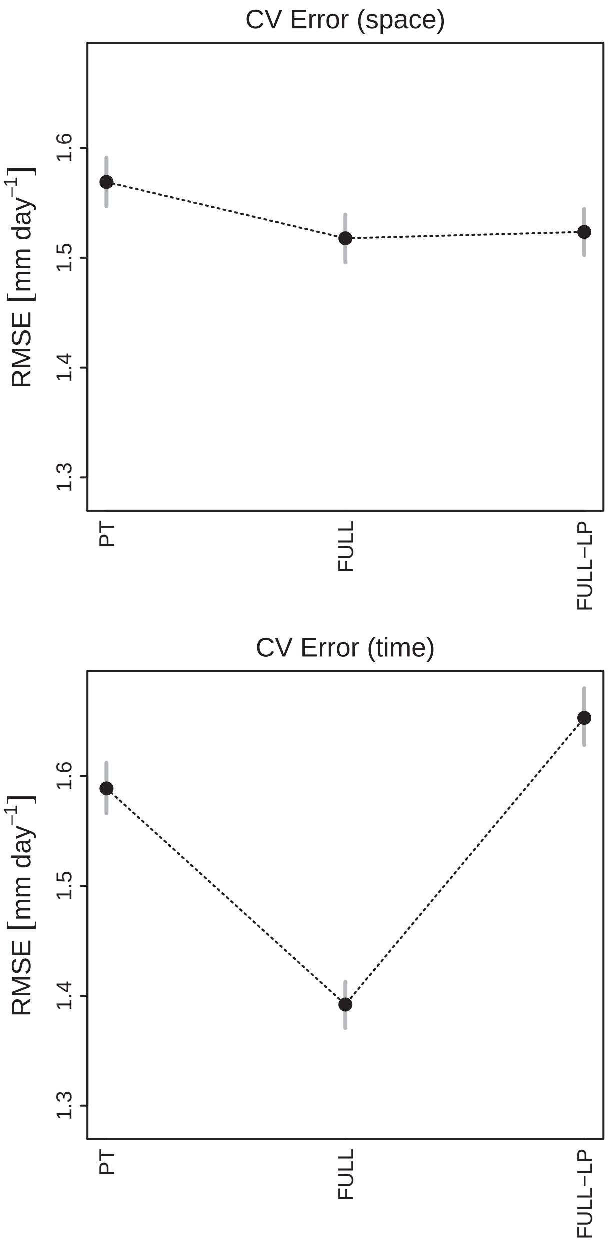

Fig. 4. Model Selection: Root mean square error (RMSE) of the three considered model setups (PT: Precipitation and Temperature forcing. FULL: Full atmospheric forcing. FULL-LP: Full atmospheric forcing and land parameters; see Section 3.1). RMSE is estimated for both the cross-validation in space and the cross validation in time (see Section 3.3.1).

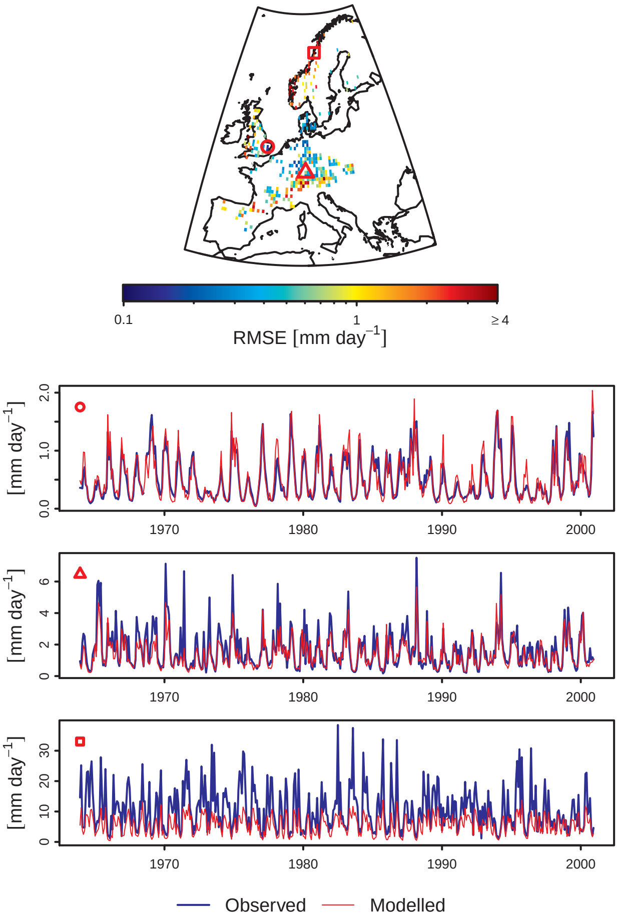

Fig. 5. Example time series: The top panel shows the RMSE of the Random Forest Model with full atmospheric forcing (Equation (2)). The symbols mark the grid cells with the lowest (circle), median (triangle) and highest (square) RMSE. The corresponding time series of observed and modelled monthly runoff are shown in the lower panels.

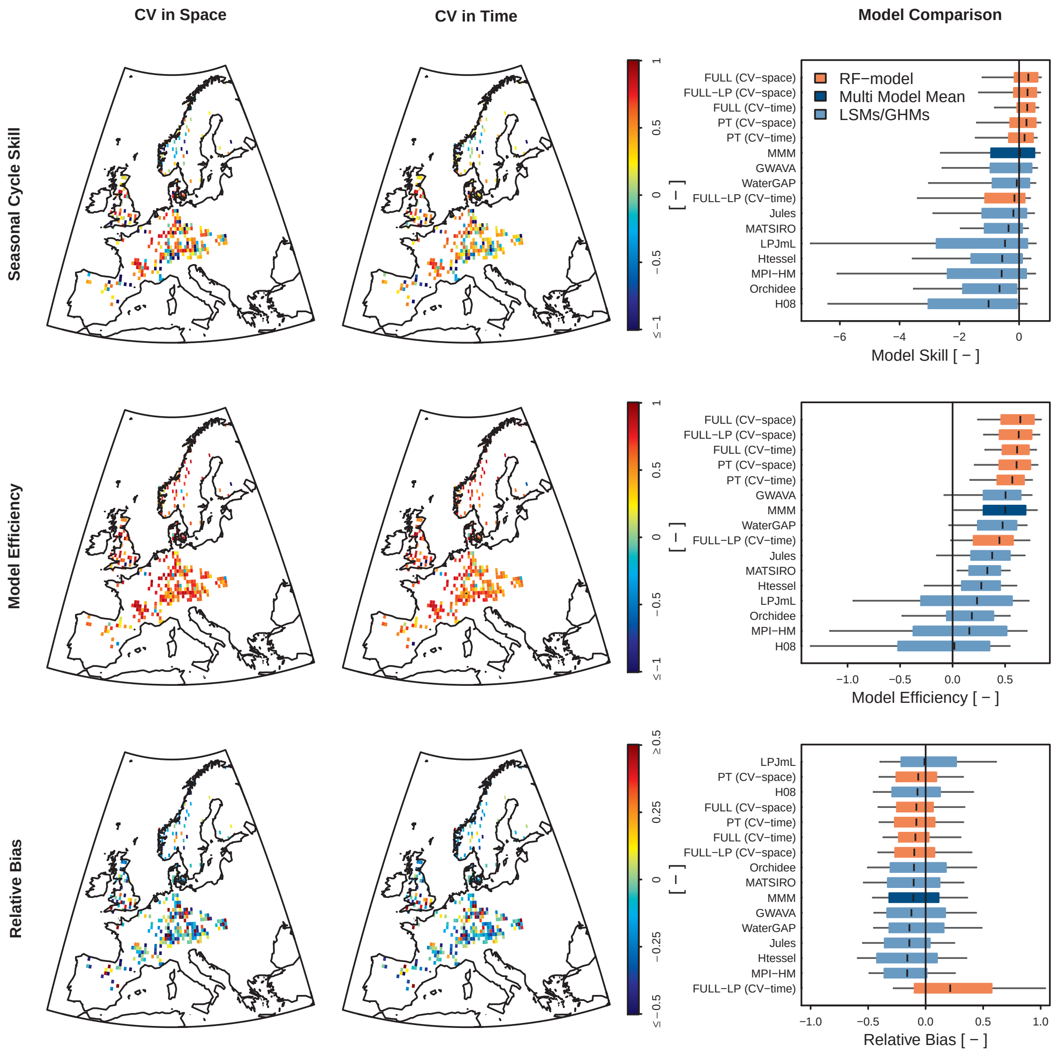

Fig. 6. Grid-cell scale validation (A): Spatial distribution of the performance of the Random Forest model with full atmospheric forcing (Equation (2)), measured with different skill scores and derived for the cross validation (CV) in time and the CV in space experiment. The boxplots allow to compare the performance distribution of all tested Random Forest models (Equations (1) to (3)) with runoff simulations from a multi model ensemble of LSMs. The individual boxes are ordered according to the median performance, such that the best performing model ranks highest.

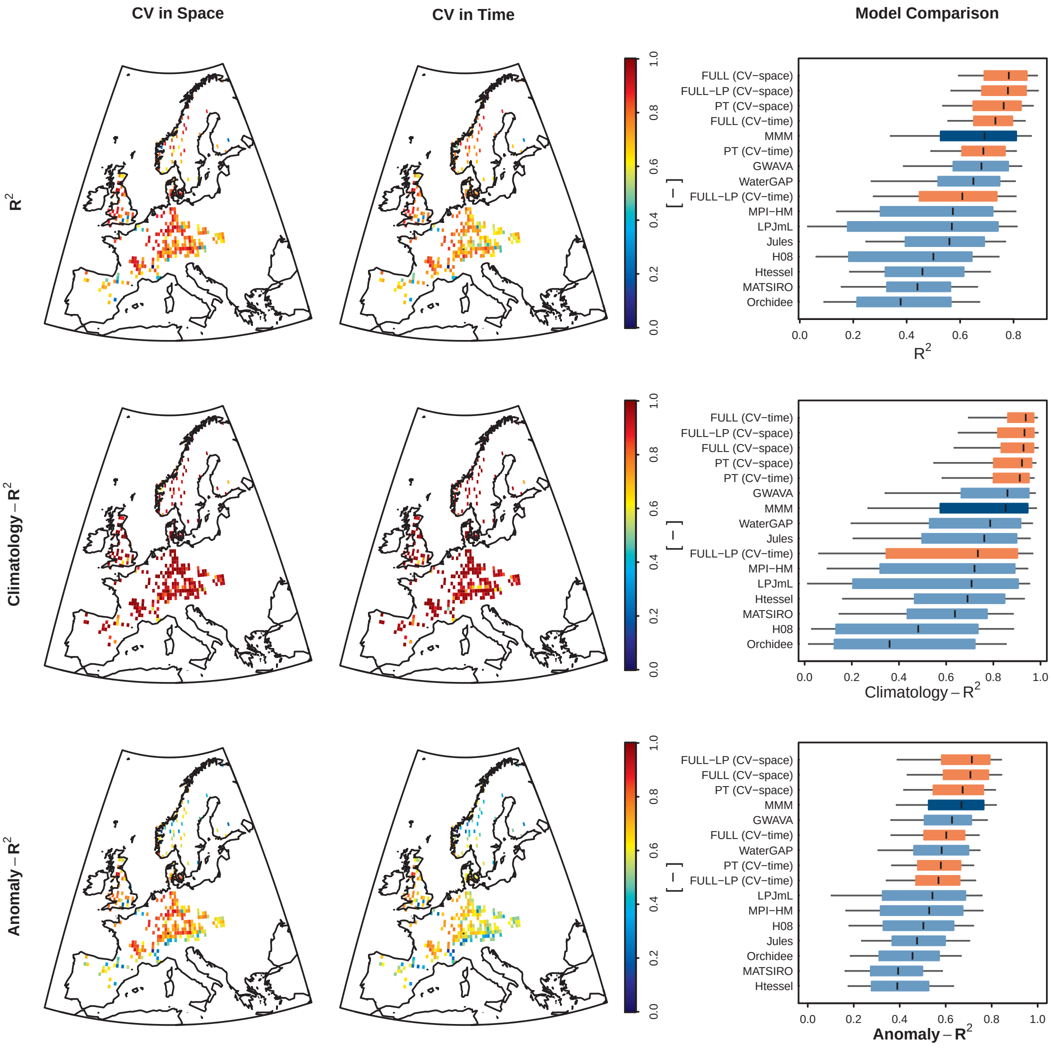

Fig. 7. Grid-cell scale validation (B): Same as Figure 6 but for different skill scores.

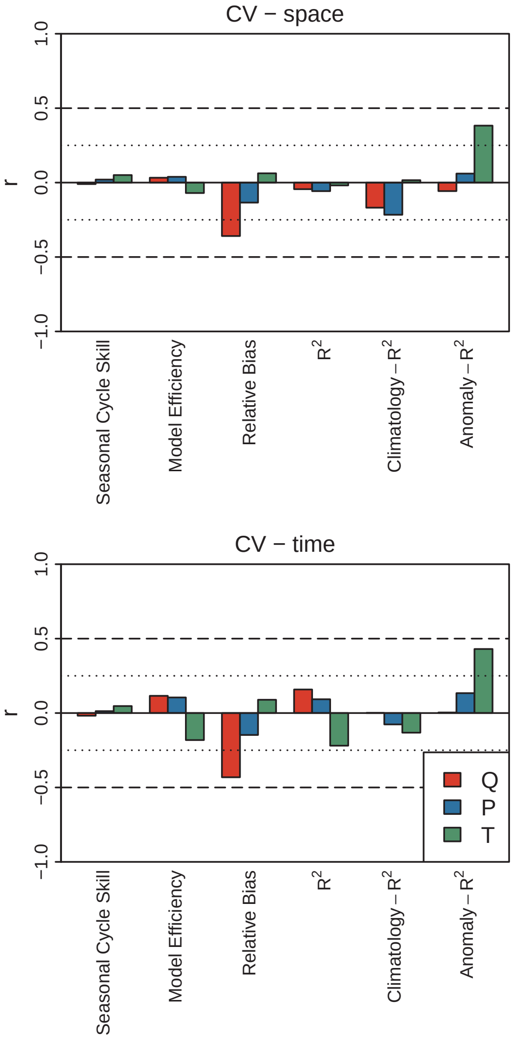

Fig. 8. Dependence of model skill on climatic conditions: Correlation between grid-cell level performance of the Random forest model with full atmospheric forcing and mean climatic conditions (Q: mean annual runoff; P: mean annual precipitation; T: mean annual temperature). Horizontal lines at $r = \pm 0.25$ and $r = \pm 0.5$ are included as a visualisation aid. Spatial patterns of the performance metrics are shown in Figures 6 and 7.

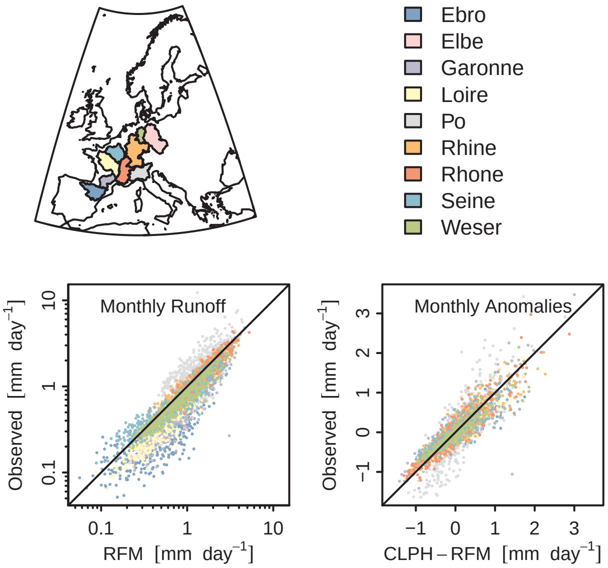

Fig. 9. Basin scale validation (A): Top, nine continental scale river basins used for model validation. Bottom, comparison between observed monthly river discharge to river discharge estimates derived from the Random Forest Model with full atmospheric forcing (left) and comparison between observed and modelled monthly discharge anomalies (right). The similarity between observed an modelled river discharge is quantified in Figure 10.

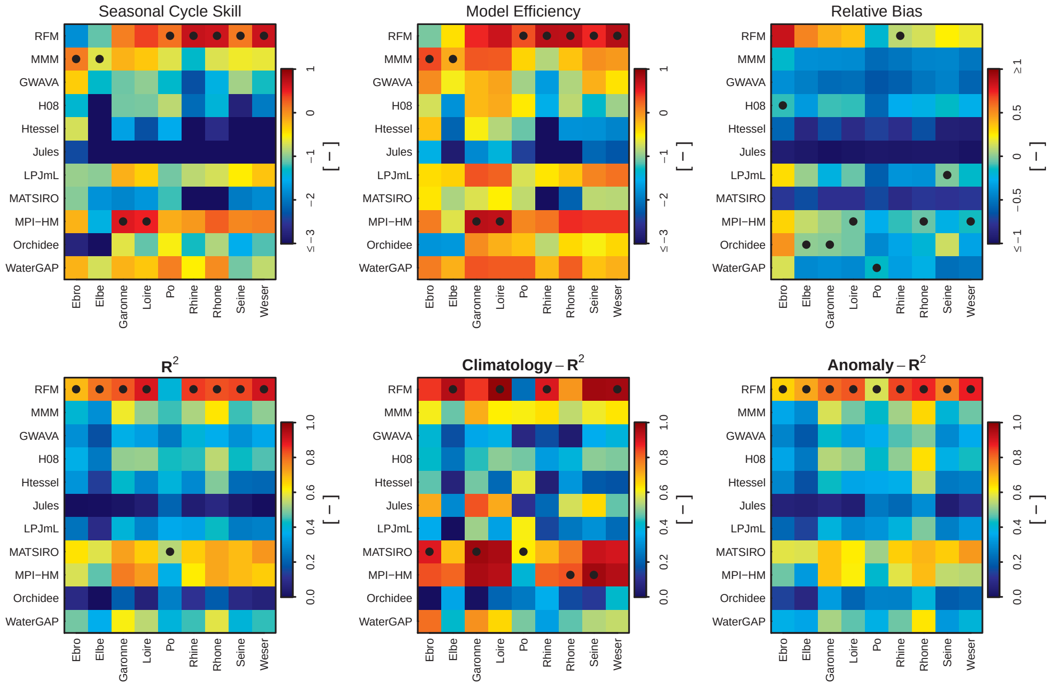

Fig. 10. Basin scale validation (B): Performance of the Random Forest Model with full atmospheric forcing compared to the performance of the considered LSMs. Model performance is assessed with respect to continental scale river discharge, quantified using six different performance metric. The best performing model for each river is marked by a dot.

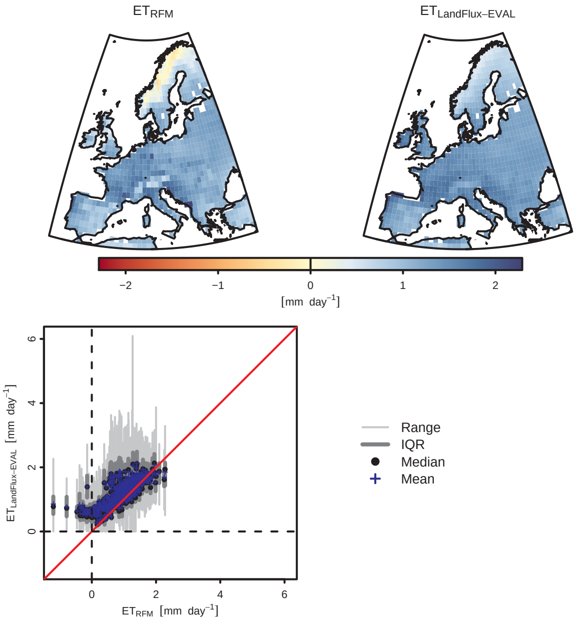

Fig. 11. Comparison of mean evapotranspiration (1989 - 1995) derived from the Random Forest Model with full atmospheric forcing and the LandFlux-EVAL synthesis product: Top left: Mean evapotranspiration computed as the mean difference between precipitation and runoff derived from the RFM. Top right: Mean evapotranspiration from the LandFlux-Eval synthesis product (Mueller et al., 2013). Bottom: Comparison of the RFM and the LandFux-EVAL estimates of mean evapotranspiration. The vertical bars denote the interquartile range (IQR) and the range of all 40 data sets entering the LandFux-EVAL product. The points and crosses indicate the median and mean evapotranspiration of the LandFlux-EVAL product.

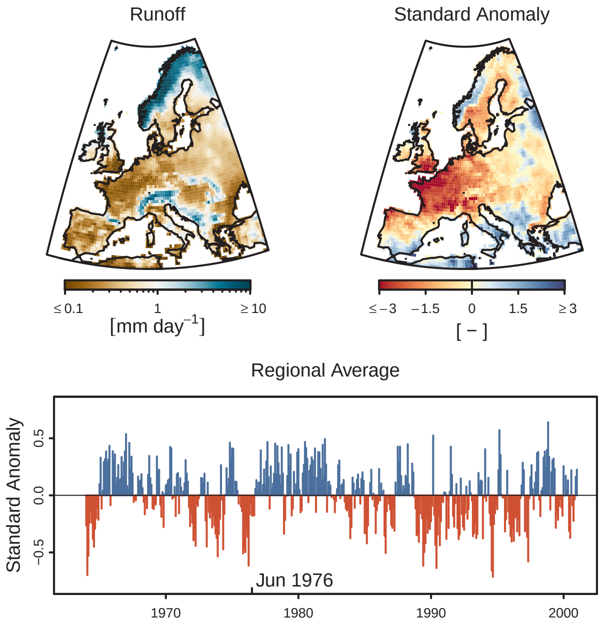

Fig. 12. The 1976 drought in Europe: The top left panel shows the monthly runoff rate in June 1976. The top right panels shows the corresponding standardised runoff anomalies. The bottom panel shows the time series of the spatial average of standardised runoff anomalies for the entire region under investigation.

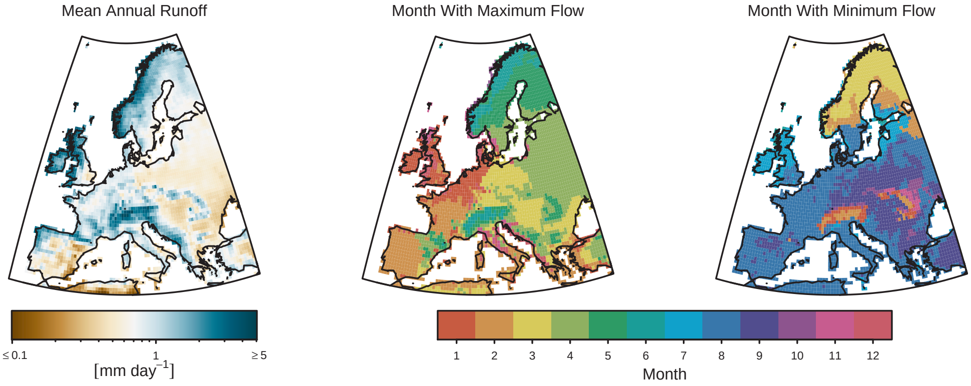

Fig. 13. European runoff climatology (1964 - 2000): Left: Long-Term mean daily runoff rates. Centre: Maximum month of the long-term mean annual cycle. Right: Minimum month of the long-term mean annual cycle.

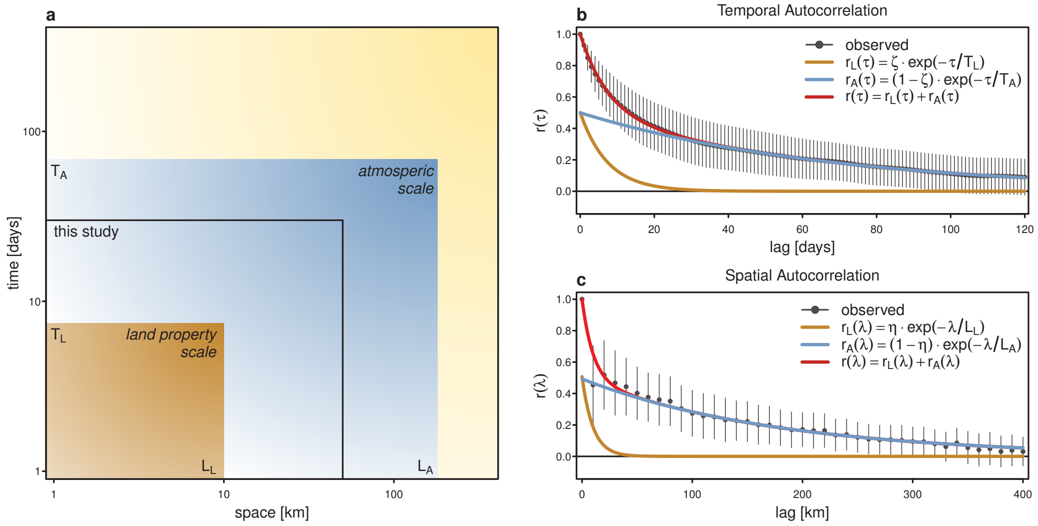

Fig. 14. Time and space scales of runoff in Europe: (**a**) Empirical results suggest that runoff in Europe has two space and time scales. A small scale (T_L: time scale; L_L: space scale), at which runoff dynamics is strongly influenced by locally varying land properties, and a large scale (T_A: time scale; L_A: space scale) at which runoff dynamics is dominated by atmospheric forcing. Both the spatial and temporal resolution of this study are located well above the scales at which land properties are expected to have a strong influence on runoff dynamics. (**b,c**) Small and large scales are estimated from observed autocorrelations of daily runoff anomalies in Europe. Vertical bars denote the standard deviation of the observed autocorrelation. See text for details.

Table 1. Median grid-cell performance of the Random Forest Model with full atmospheric forcing (Equation (2))

	CV in Space	CV in Time
S_{seas}	0.31	0.27
MEf	0.64	0.61
BIAS	-0.08	-0.09
R^2	0.78	0.73
R^2_{clim}	0.93	0.94
R^2_{ano}	0.71	0.60

Table 2. Temporal and spatial scales of daily runoff in Europe: Estimate, standard error and p-value (t-test) of the scaling models (equations (8) and (9)) fitted to observed temporal and spatial correlation functions using nonlinear least squares regression. Note, that the lower limit of L_L was set to the resolution of the empirical spatial correlation function (10 km).

		Temporal			Spatial	
	ζ [-]	T_L [days]	T_A [days]	η [-]	L_L [km]	L_A [km]
Estimate	0.50	7.4	68.3	0.51	$\leq$10	180.5
Standard Error	3.8×10^{-3}	0.1	0.6	0.04	2.9	19.6
p - value	<0.001	<0.001	<0.001	<0.001	0.002	<0.001