Revisiting the potential to narrow model uncertainty in the projections of Arctic runoff

Emma Dutot[1] and Hervé Douville[1]

[1]Météo-France

April 11, 2023

Abstract

Despite multiple advances in the understanding of the water cycle intensification in a warmer climate, climate models still diverge in their hydrological projections. Here we constrain annual runoff projections over individual and aggregated Arctic river basins. For this purpose, we use two ensembles of global climate models and two statistical methods: a regression scheme assuming similar runoff sensitivities at interannual versus climate change timescales, and a Bayesian method where models are used to derive a posterior runoff response conditional to historical observations. While both techniques are shown to narrow model uncertainties, more or less substantially depending on rivers, the Bayesian method is less sensitive to the choice of the model ensemble and is more skilful when tested with synthetic observations. It has also been applied over the whole Arctic watershed, showing so far a limited narrowing of the inter-model spread, but its skill will further improve with increasing climate change.

Revisiting the potential to narrow model uncertainty in the projections of Arctic runoff

Emma Dutot[1] and Hervé Douville[2*]

[1] Météo-France, Lyon, France

[2] Centre National de Recherches Météorologiques, Université de Toulouse, Météo-France, CNRS, Toulouse, France

* Corresponding author : herve.douville@meteo.fr

Abstract

Despite multiple advances in the understanding of the water cycle intensification in a warmer climate, climate models still diverge in their hydrological projections. Here we constrain annual runoff projections over individual and aggregated Arctic river basins. For this purpose, we use two ensembles of global climate models and two statistical methods: a regression scheme assuming similar runoff sensitivities at interannual versus climate change timescales, and a Bayesian method where models are used to derive a posterior runoff response conditional to historical observations. While both techniques are shown to narrow model uncertainties, more or less substantially depending on rivers, the Bayesian method is less sensitive to the choice of the model ensemble and is more skilful when tested with synthetic observations. It has also been applied over the whole Arctic watershed, showing so far a limited narrowing of the inter-model spread, but its skill will further improve with increasing climate change.

Plain language summary:

Despite considerable progress in understanding the intensification of the water cycle in response to global warming, projections of changes in the various components of the continental water balance remain highly model-dependent. This is particularly the case for the evolution of runoff, which remains very sensitive to the way atmospheric and continental processes are represented. Here we evaluate and compare two statistical methods to constrain climate projections of annual runoff at high latitudes via available or synthetic observations. Our Bayesian method, based on historical observations, is found to be more robust than a previous method based on observed interannual climate variability. It underlines the importance of having reliable flow reconstructions to constrain runoff projections and a careful use of more empirical emergent constraints which can lead to overconfident projections.

Key points :

- Runoff sensitivity to temperature depends on the considered timescale.
- Observational constraints based on interannual variability can thus lead to overconfident projections.
- Observational constraints based on historical trends can be effective if such trends are driven by human emissions of greenhouse gasses.

1. Introduction

According to the latest IPCC report, hydrological projections at the regional scale are still very uncertain despite a robust theoretical understanding of the water cycle intensification at the global scale (AR6 WG1, Douville et al., 2021). More specifically, there is low confidence in the sign and magnitude of projected changes in global land runoff in all illustrative Shared Socio-economic Pathway (SSP) scenarios scrutinized in the sixth phase of the Coupled Model Intercomparison Project (CMIP6). There are however some regional exceptions, especially in the northern high latitudes where there is high confidence that projected increases in precipitation amount and intensity will be associated with increased runoff throughout the 21st century.

Much of the uncertainty in hydrological projections arises from global climate model (GCM) uncertainty rather than emission scenario uncertainty, while internal climate variability also plays a key role in the next two decades (Lehner et al., 2020 ; Douville et al., 2021). Moreover, CMIP6 often shows larger model uncertainties than CMIP5 (Lehner et al., 2020), which may be partly due to a wider range of climate sensitivity (Hausfather et al., 2022). Yet, this explanation may not hold for runoff whose regional response does not scale with global warming across multiple models given the influence of other potential drivers, including model-dependent changes in large-scale circulation (Douville et al., 2022; Elbaum et al., 2022).

A recent global evaluation of runoff simulations, from both CMIP6 GCMs and off-line global hydrological models, suggests that models generally perform better in non-cold environments and that the multi-model ensemble average is not systematically an effective way to reduce bias compared to individual models (Hou et al., 2023). These results further support previous findings based on CMIP5 GCMs that urged caution in the direct use of climate model runoff for hydrological and water security applications (Lehner et al., 2019). Regional climate models at higher spatial resolution are extremely useful to project runoff in small watersheds but do not necessarily provide more reliable projections in large-scale river basins since they are strongly constrained by their driving GCMs. Non-linearities in the runoff response to global warming may also challenge the use of simple pattern-scaling techniques to produce reliable hydrological projections at the

regional scale (Zhang et al., 2018 ; Douville et al., 2021).

In response to these difficulties, an increasingly popular approach is to use emergent constraints to link future changes to present-day or recent climate features across multiple GCMs. Unfortunately, most emergent constraints that have been proposed to narrow uncertainties in CMIP5 projections generally do not perform as well for CMIP6 models (Sanderson et al., 2021). These results support previous studies concluding that emergent constraints should be based on an independently verifiable physical mechanism (e.g., Hall et al., 2006) and should be considered with great caution when there is no such dominant single mechanism but, as often, a plurality of poorly-constrained mechanisms due to possible gaps in both models and observations. Emergent constraints have thus the potential to produce overconfident projections, especially when processes are represented in a common way throughout the GCM ensemble (Sanderson et al., 2021). Moreover, they sometimes assume a time-invariance of climate feedbacks so that the mechanisms that dominate future climate changes are considered to be similar and observable on shorter timescales such as the annual cycle (Hall et al., 2006) or interannual variability (Douville et al., 2006 ; Lehner et al., 2019). This is a strong hypothesis given the growing evidence that climate feedbacks may not only vary across timescales (e.g., Dessler and Forster, 2018) but may also be time-dependent on climate change timescales (e.g., Andrews et al., 2022).

Only few studies have quantified regional runoff projections in CMIP models, generally demonstrating a wide inter-model spread but a strong consistency in the ensemble mean model behavior across emissions scenarios or model generations (e.g., Tang et al., 2012 ; Zhang et al., 2018 ; Giuntoli et al., 2018). Even less studies have tried to go beyond the "one model one vote" approach in order to assess more carefully the regional runoff sensitivity (e.g., Lehner et al., 2019 ; hereafter L19). Unlike in off-line global hydrological models, runoff sensitivity is rarely explicitly tuned in GCMs, thus offering an opportunity to develop an observational constraint unaffected by model calibration efforts. In L19, there are however two key sensitivities for which the time-invariance assumption has been made: the precipitation sensitivity of runoff (sometimes termed ‘runoff elasticity’ ; e.g., Tang and Lettenmaier, 2012), defined as the percent change in runoff induced by a 1% change in precipitation ($\Delta Q/\Delta P$); and the temperature sensitivity of runoff, defined as the percent change in runoff for a 1°C change in temperature ($\Delta Q/\Delta T$). Such metrics can be diagnosed

from observations but their relevance is not guaranteed when it comes to constrain runoff projections across the whole 21st century.

In the present study, we illustrate this issue by revisiting L19 and downrating the potential to constrain the runoff response to climate change from the observed interannual variability. As in L19, the focus is on the northern high latitudes (although here including both North America and northern Eurasia) where a significant fraction of runoff originates as snowmelt and may be thus particularly sensitive to global warming (e.g., Li et al, 2017). After a brief description of the various datasets and statistical methods (Section 2), the limitations of the L19 method are first shown by using two definitions of runoff and two generations of GCMs in a high-emission scenario over four northern high-latitude river basins (Section 3). Moreover, the results of the L19 method are compared to those obtained with an alternative statistical technique recently developed at CNRM, the so-called Krigging for Climate Change technique (hereafter KCC; Ribes et al., 2020 ; Qasmi and Ribes, 2022). For the sake of a fair comparison, a pseudo-observation setting is also used in which one GCM is discarded from the CMIP ensemble and used to provide both historical and future data. Finally (Section 4), the reasons for the better behavior of the KCC method are briefly discussed and the method is also applied to constrain runoff projections over the aggregated "Arctic" basin (i.e., river basins whose outlet is in the Arctic ocean or neighboring seas).

2. Data and methods

We work with three main variables, total precipitation (P), near-surface air temperature (T) and total runoff (R), that have been averaged over large river basin and throughout the hydrological year (from October of the year N to September of the year N+1, as in L19). The focus is on the northern high-latitudes where there is an amplification of global warming, a human-induced increase in annual mean precipitation (Douville et al., 2021) and an expected climate change signal in the observed runoff reconstructions.

2.1 Observed and simulated data

The main observed monthly datasets used are precipitation from GPCC (1901-2018), near-surface air temperature from CRU_TS4 (1901-2019) or HadCRUT5 (1850-2020), and runoff from GRUN (1902-2013). Note that GRUN does not provide "real" observations, but is based on a machine learning method that estimates runoff from T and P observations (Ghiggi et al., 2019). The model is trained with monthly observations at the scale of relatively small catchment areas and then is applied on gridded data from reanalysis.

For simulated data, historical and future simulations (1850-2100) are taken from the last two phases of the Coupled Model Intercomparison Project (CMIP5 and CMIP6). We worked with the comparable although slightly different high-emission scenarios (RCP8.5 and SSP5-8.5 for CMIP5 and CMIP6 respectively) in order to maximize the signal to noise ratio. We distinguished several multi-model ensembles: ensembles with only one realization (data from the first available member of each model, we named these ensembles rcp81 and ssp51 respectively); a CMIP6 ensemble with at least 3 members but less models although we only used the average of all available realizations for each model (named ssp50), and a few mono-model ensembles (ssp5x) with 25 individual members.

2.2 Statistical methods used to constrain basin-scale runoff projections

The CMIP projections of 12-month runoff averaged from October to next September (the so-called water year) have been constrained with two different statistical methods. The first method is the L19 method depicted in Lehner et al. (2019) and based on runoff predictions from a multiple linear regression whose coefficients are constrained with the observed interannual variability. In the end, we can compute the basin-wide average runoff anomaly $\Delta R = a_0*\Delta P + b_0*\Delta T$, where a_0 and b_0 are the regression coefficients computed from observations (observed interannual variability) over the common 1902-2013 period. This period will be also referred to as the training period when L19 will be evaluated with model outputs as synthetic observations.

The second method (KCC) was developed by Ribes et al. (2021) and Qasmi and Ribes (2022) and is based

on Bayesian statistics where a *prior* distribution, $\pi(x)$, of the forced response to anthropogenic forcings is first derived from raw model outputs and then constrained directly with observations of one or more variables (here both observed global mean surface temperature and reconstructed basin-wide average runoff). The *prior* is estimated using a Generalized Additive Model (GAM, assuming the additivity of the model responses to individual forcings) and a simple Energy Budget Model (EBM, allowing us to diagnose the runoff response to volcanic eruptions; for more details, see supplementary materials from Ribes et al., 2021). For the sake of simplicity, this *prior* is assumed to follow a normal distribution and thus only needs an estimate of the ensemble mean and spread. Next, observations y are used to derive a *posterior* distribution (after constraint). We suppose that observations can be described as: $y=\mathcal{H}*x+\varepsilon$, where $\mathcal{H}$ is a pseudo-observation operator allowing to extract the part of x observed in y and ε corresponds to internal variability and observational errors (if available). Since $\pi(x)$ and ε are assumed to follow normal distributions, the *posterior* can be easily derived using the Gaussian conditioning theorem (for more information, see supplementary materials from Ribes et al., 2021).

These two statistical methods are thus very different. L19 is using the observed internal variability to constrain the regression coefficient of a multiple regression scheme aimed at predicting the long-term evolution of runoff from the evolution of the basin-wide average precipitation and near-surface temperature. In contrast, KCC is using directly the observed evolution of runoff and/or of global warming over the historical period to constrain the forced response of the simulated runoff with a Bayesian statistical method. Given the use of interannual runoff fluctuations to fit the regression scheme, L19 needs individual realizations of each model. For the sake of simplicity and of a fair comparison between the two methods, we thus start to work with the ssp51 and rcp81 multi-model ensembles. Yet, KCC focuses on the forced response and is thus expected to provide even better results when using as many realizations as possible for each model, hence also the use of the ssp50 ensemble based on the average of all members for each model.

2.3. Pseudo-observations and related scores

To assess and compare the reliability of the L19 and KCC methods, we decided to work with

pseudo-observations (as done in Ribes et al., 2021). The idea is that we extract one of the models and we use it as synthetic observations of both past and future variations of T, P and R. In the end, after constraining our remaining projections using the early pseudo-record (1901-2013), we can check if they are getting closer to our R pseudo-observations over the end of the 21st century compared to the raw (i.e., unconstrained) runoff projections. Each model provides successively the reference pseudo-observations, so that we can compute several probabilistic scores to assess the quality and reliability of our constrained projections. Beyond the inter-model spread or the multi-model 90% confidence interval, two main scores were computed in this study : the coverage probability (CP) and the CRPS change.

- Coverage probability (CP): for each iteration, we check that the future pseudo-observations for the 2081-2100 period (water-year mean anomalies over this period) are in the 5-95% confidence interval of the constrained projections. In the end, we get a CP percentage corresponding to the number of reference models (POBS) in this confidence interval, which should be as close to 90% as possible.
- CRPS is a probabilistic score quantifying the distance between an observation (here corresponding to "pseudo-observed" 20-water-year mean runoff anomalies) and the distribution of predicted anomalies over the same 2081-2100 period. The closer the observation is to the predicted distribution, the closer the CRPS is to 0. We compared the CRPS computed with the model distribution before versus after constraint. Then we computed the relative change in CRPS: a negative change will thus mean that observation and predictions are closer after constraint (i.e., improved distribution). For each basin, we averaged this CRPS change after using successively each model as pseudo-observations.

3. Results

Our main purpose here is first to constrain the ssp51 and rcp81 ensembles of runoff projections with observations, using two different statistical methods. We will compare the robustness and reliability of the two methods by first checking the sensitivity to the chosen model ensemble (CMIP6 vs CMIP5), and then computing statistical scores using the idealized pseudo-observation framework.

3.1 Assessing the robustness of the L19 method

In Lehner et al., 2019, the L19 method was applied to CMIP5 models and runoff was estimated as P-E (precipitation - evapotranspiration), thus assuming no change in water storage. In the present study, we work with CMIP6 models and genuine simulated runoff outputs. In agreement with L19 (cf. their Fig.1), Fig.S1 in the supplementary information (SI) shows that runoff anomalies computed with a multiple linear regression (using both P and T predictors) performs better than a simple regression (using P only) to predict runoff anomalies over the training period (1901-2013). Note that other regressions could be also tested but would not be as effective or based on less reliable predictors in the instrumental record. We thus choose to work with the same basin-wide average P and T predictors as in L19. A similar figure but using P-E instead of R (as in Lehner et al., 2019) is shown in SI (Fig.S2). Not surprisingly, regressions show an improved skill when using P-E rather than R. Thus, using P-E as an approximation of R can lead to overconfident projections when using the L19 method. Thereafter we will work with the R variable averaged throughout the water year, for which river the observed discharge provides reasonable estimates.

The L19 method apparently succeeds in narrowing uncertainty in runoff projections. Indeed, the 5-95% confidence interval is reduced after the L19 constraint (black and green line, Fig.1) compared to the simulated and raw predicted anomalies at the end of the 21st century. Overall, even when uncertainties in the observed regression coefficients are accounted for (green thin line), we obtain a systematic reduction of the 90% confidence interval. Similar qualitative results are obtained when using P-E as a surrogate of runoff (Fig.S3). Yet, our results are basin-dependent and also sensitive to the choice of the multi-model ensemble as shown by the comparison between CMIP6 (Fig.1) and CMIP5 (Fig.S4).

To sum up, the L19 method seems to succeed in narrowing the uncertainties in runoff projections over the four selected northern high-latitude river basins : Columbia, Kolyma, Lena and Mackenzie. However, these results rely on severe assumptions that have not been tested properly and may lead to overconfident projections. This is the reason why we will further evaluate the reliability of the L19 method by using synthetic observations of both past and future climate in Section 3.3.

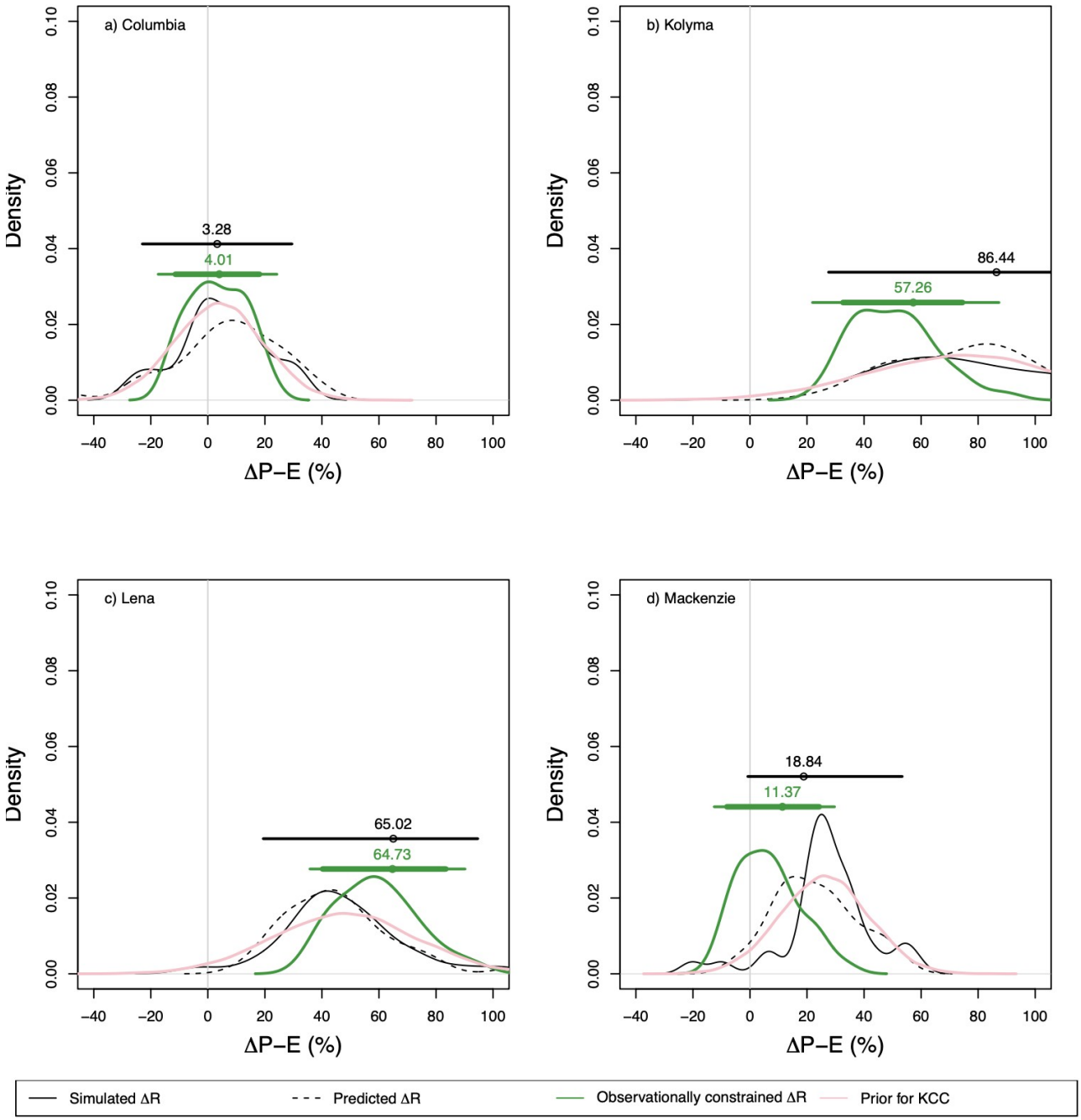

Figure 1: Constrained versus unconstrained distributions of water-year mean relative runoff anomalies (%) from the CMIP6 model ensemble under the SSP5-8.5 high-emission scenario : a) Columbia, b) Kolyma, c) Lena, and d) Mackenzie. Kernel density functions of runoff anomalies as simulated by CMIP6 models and as predicted and constrained by the L19 method (runoff anomalies are averaged over the 2081-2100 period). The KCC prior distribution is also shown ($\pi(x)$ in red) in order to highlight that the normal distribution assumption has a significant impact on the prior distribution given the limited size of the CMIP6 ensemble. Predicted runoff anomalies are those computed from the L19 multiple regression but without constraining the regression coefficients with observations. The black and green horizontal lines

represent the 5-95% confidence interval of the simulated and the L19 constrained projections respectively. The green confidence interval includes uncertainties in the regression coefficients as estimated with a bootstrap method (green thin line).

3.2 Assessing the robustness of the KCC method

The KCC method was also applied to constrain both CMIP6 and CMIP5 annual runoff projections. Two metrics were used as observational constraints: the global mean near-surface air temperature from HadCRUT5 (including an estimate of observational errors) and/or the basin-wide average runoff from GRUN (without error estimates and based on reconstructed runoff from observations rather than genuine river discharge measurements). For the four selected river basins, the CMIP6 inter-model spread of the distribution is reduced by around 10 to 35% in 2100 after applying the KCC observational constraint (Fig.2). Similar results are obtained with the CMIP5 models (Fig.S5). The limited narrowing of model uncertainty, compared to the use of L19 and in particular for the Columbia river also considered in L19, is consistent with the lack of an obvious long-term trend in the GRUN reconstruction. When comparing the constrained projections between ssp51 (Fig.2) and ssp50 (with multiple members for each model, Fig.S6), similar results are found thereby suggesting that internal climate variability is not responsible for the apparently lower performance of KCC.

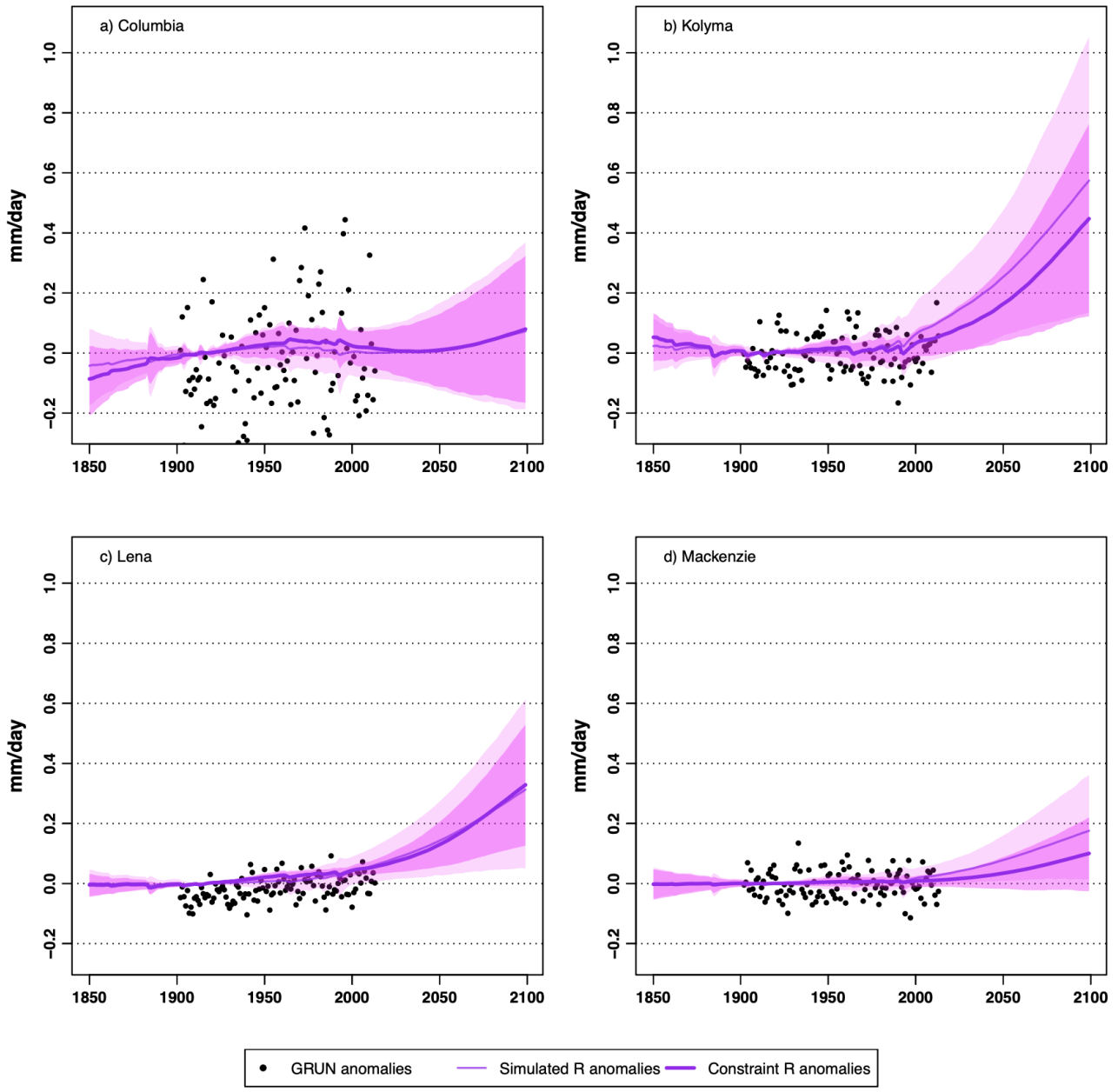

Figure 2: Constrained versus unconstrained water-year runoff anomalies (mm/day) using the ssp51 ensemble (a single realization of each CMIP6 model under the SSP5-8.5 scenario). Black dots correspond to the GRUN water-year runoff anomalies. Both GRUN runoff and HadCRUT5 GSAT observations are used to constrain the simulated runoff anomalies. The thick lines denote the best estimate of each distribution (i.e., the ensemble mean) while shadings denote the corresponding 5-95% confidence intervals.

The KCC method was also applied using only either the global mean surface temperature observations (GSAT from the HadCRUT5 dataset, consisting of 200 members to account for observational uncertainties) or the basin-wide runoff GRUN reconstruction (assuming no observational uncertainty). When using only

GSAT observations, the inter-model spread is not much reduced, thereby supporting the claim that regional hydrological changes are not heavily constrained by global warming (Douville et al., 2022). This finding highlights the need for monitoring river discharge and providing estimates of water withdrawals versus natural runoff, including observational uncertainties, for better constraining runoff projections.

To sum up, KCC was also applied to Columbia, Kolyma, Lena and Mackenzie and led to a weaker (by 10% to 35%) but more robust (CMIP5 versus CMIP6) narrowing of model uncertainty in the highest emissions scenarios and using a single member for each model. Yet, the performance of KCC is expected to improve with the emerging influence of climate change in the observed time series, in contrast to L19 where longer observed time series will only have a limited impact on the observed regression coefficients dominated by interannual variability. Moreover, a fair comparison between KCC and L19 necessitates a different setting where physically-consistent pseudo-observations are derived from a randomly-chosen climate model in order to assess the constrained projections based on the other models.

3.3 Validation of these results with pseudo-observations

We showed that both L19 and KCC methods have the potential to reduce model uncertainty on projected runoff anomalies in a warmer climate. Yet, we do not know so far the extent to which these results are reliable. As in Ribes et al.(2021), we now use a pseudo-observation framework and compute three probabilistic scores (CP, CRPS change and reduction of spread) in order to assess the robustness and reliability of our methods (see section 2.2 for details). Results are compared over the four selected watersheds in order to get an overall picture of the relative performance of the L19 and KCC methods (table S1, S2 and S3). As a reminder, and beyond the expected reduction in the inter-model spread, CP has to be as near to 90% as possible and CRPS has to be as low (close to zero) as possible after constraining the projections. Therefore, the change in CRPS and spread have to be negative for the constraint to be reliable and effective.

Although leading to a fairly similar systematic reduction of uncertainty, the L19 and KCC methods show

contrasted scores in terms of both CP and CRPS. For the L19 method, CP is far from 90% and changes in CRPS are systematically positive, thus providing clear evidence that the method leads to overconfident projections, likely due to a timescale-dependent temperature effect on runoff (cf. Fig.S7 and S8). For the KCC method, CP is above 80% for three out of four basins (table S3) and CRPS changes are either negative or slightly positive. While these results suggest a cautious use of both methods, they clearly emphasize the superiority of the Bayesian KCC technique compared to the more empirical L19 emergent constraint.

Regarding the KCC method, scores are generally (and as expected) slightly improved if one uses several rather than just one realization to assess the forced response of each model (ssp50, compared to ssp51; better spread reduction, overall better change in CRPS but slightly worse CP). Yet, they are still basin-dependent and better when using the local (GRUN) constraint only (not shown here). This finding highlights again that the observed global historical warming has little positive influence on the constrained runoff and further supports the cautious statement of Douville et al. (2022) regarding the scalability of hydrological changes with global warming across multiple models.

Note that we also tried to constrain the basin-scale projections of precipitation and temperature using KCC, before applying the L19 method (multiple linear regression with observationally-constrained regression coefficients) to constrain the projections of runoff. However, this combined method led to even more overconfidence than applying the L19 method only when tested with pseudo-observations, so that the results were not included in this article.

To sum up, scores are overall better when using the KCC method to constrain runoff projections with pseudo-observations compared to the use of the more empirical L19 method. The overconfidence of L19-constrained runoff projections arises primarily from the contrasted runoff sensitivity to temperature and precipitation at climate change versus interannual timescale (cf. Fig.S7 and S8).

4. Discussion

Two statistical methods were applied with available observations, and compared in a more idealized setting, to reduce the uncertainty about projected northern-latitude runoff changes throughout the 21st century at the basin-scale the L19 method (Lehner et al., 2019) and the KCC method (Ribes et al., 2021). After using pseudo-observations to compute statistical scores and assess the reliability of our results, we found that the L19 method is systematically overconfident, which means that there is a greater than 10% probability for the actual runoff evolution to lie outside the constrained 5-95% confidence interval. Generally speaking, the KCC method leads to more reliability and better scores, though variable from one basin to another depending on the observed trend in the GRUN natural runoff reconstructions (not accounting for a direct human influence through water withdrawals and river management). The overconfidence of the L19 results is primarily due to a wrong hypothesis regarding the runoff sensitivity to temperature and precipitation at different timescales (e.g., Zhang et al., 2022). Our results thus illustrate the danger of using emergent constraints without having first tested them in an idealized context (Sanderson et al., 2021). Furthermore, they highlight the potential of more direct observational constraints whose power should increase with further climate change and stronger signal to noise ratio in observed runoff time series.

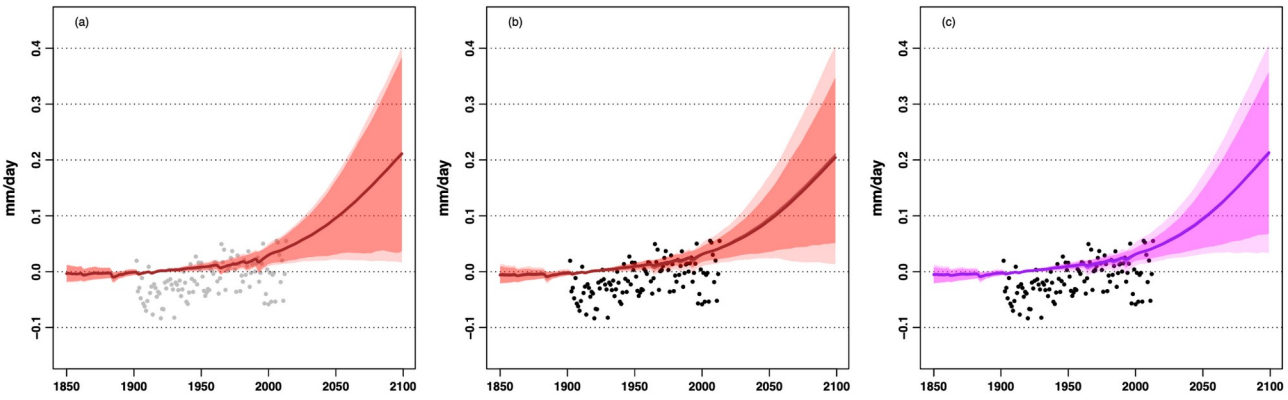

Figure 3 : Constrained versus unconstrained aggregated Arctic water-year runoff anomalies (mm/day) using the ssp50 ensemble a) KCC constraint using only the HadCRUT5 GSAT observations. **b**) KCC constraint using only the GRUN runoff reconstruction. **c**) KCC constraint using both HadCRUT5 and GRUN observational constraints. Black dots correspond to the GRUN water-year anomalies; they are colored in grey (rather than black) when GRUN is not used to constrain the projections. The thick lines (here

superimposed) denote the best estimate of each distribution (i.e., the ensemble mean) while shadings denote the corresponding 5-95% confidence intervals.

To conclude this study, we have thus decided to apply the KCC method to the aggregated "Arctic" basin by merging all watersheds whose outlets are in the Arctic Ocean and nearby seas. Results of the individual and combined constraints are shown Fig.3. Overall, the constraint using both GSAT and runoff observations (GRUN and HadCRUT5) allows KCC to reduce the CMIP6 inter-model spread in Arctic runoff by 22% (Fig.3.c) in 2100 (even more in the early 21st century). As expected when increasing the signal to noise ratio through spatial aggregation, this combined constraint is more efficient than the individual ones using either HadCRUT5 or GRUN reconstructions. Scores with pseudo-observations for the whole Arctic basin (not shown) were also computed. They are quite promising and suggest even better results after one more decade of observations. Our study thus emphasizes the need of more reliable and routinely updated runoff observations (compared to GRUN for instance) to constrain model projections that, unfortunately, do not show a spread reduction from one model generation to the next despite the sustained efforts of global modeling centers to improve and evaluate such models.

Data Availability Statement

We did not use new data in the present study. The CMIP5 and CMIP6 monthly mean model outputs are available on the ESGF archive at https://esgf-node.llnl.gov/, GRUN reconstruction of monthly runoff are available at https://www.researchcollection.ethz.ch/handle/20.500.11850/324386, HadCRUT5 global mean surface temperature is available at https://www.metoffice.gov.uk/hadobs/hadcrut5/.

Code Availability Statement

The KCC statistical package for observational constraint is available on gitlab at https://gitlab.com/saidqasmi/KCC. Other codes for CMIP5 and CMIP6 data curation and visualization are based on the CliMAF package available at: https://climaf.readthedocs.io/en.

References :

Andrews, T., Bodas-Salcedo, A., Gregory, J. M., Dong, Y., Armour, K. C., Paynter, D., et al. (2022). On the effect of historical SST patterns on radiative feedback. Journal of Geophysical Research: Atmospheres, 127, e2022JD036675. https://doi.org/10.1029/2022JD036675

Dessler, A. E., & Forster, P. M. (2018). An estimate of equilibrium climate sensitivity from interannual variability. *Journal of Geophysical Research:Atmospheres*,123, 8634–8645. https://doi.org/10.1029/2018JD02848

Douville, H., Salas-y-Mélia, D., Tyteca, S. (2006) On the tropical origin of uncertainties in the global land precipitation response to global warming. *Clim. Dyn.,* 26, 367–385, doi :10.1007/s00382-005-0088-2

Douville, H., Raghavan, K., Renwick, J., Allan, R.P., Arias, P.A., Barlow, M. et al. (2021) Water Cycle Changes. In Climate Change 2021: The Physical Science Basis. *Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 1055–1210, https://doi.org/10.1017/9781009157896.010

Douville, H., Allan, R.P., Arias, P.A., Betts, R.A., Caretta, M.A., Cherchi, A., et al. (2022) Water remains a blind spot in climate change policies. *PLOS Water*, 1(12): e0000058. https://doi.org/10.1371/journal.pwat.0000058

Douville, H., Qasmi, S., Ribes, A., Bock, O. (2022) Global warming at near-constant relative humidity further supported by recent in situ observations. *Communications Earth & Environment*, 3, 237, https://doi.org/10.1038/s43247-022-00561-z

Elbaum, E., Garfinkel, C.I., Adam, O., Morin, E., Rostkier-Edelstein, D., Dayan, U. (2022) Uncertainty in projected changes in precipitation minus evaporation: dominant role of dynamic circulation changes and weak role for thermodynamic changes. *Geophys. Res. Lett.,* https://doi.org/10.1029/2022GL097725

Ghiggi, G., Humphrey, V., Seneviratne, S. I. and Gudmundsson, L. (2019) GRUN : an observation-based global gridded runoff dataset from 1902 to 2014. *Earth System Science Data*, 11, 1655–1674, https://doi.org/10.5194/essd-11-1655-2019

Giuntoli, I., Villarini, G., Prudhomme, C., Hannah, D.M. (2018) Uncertainties in projected runoff over the conterminous
United States. *Climatic Change*, 150, 149–162, https://doi.org/10.1007/s10584-018-2280-5

Hall, A., and X. Qu (2006), Using the present-day seasonal cycle to constrain climate sensitivity: A case study of snow albedo feedback, *Geophys. Res. Lett.*, 33, 1550–1568, doi:10.1029/2005GL025127

Hausfather, Z., Marvel, K., Schmidt, GA., Nielsen-Gammon, JW., Zelinka, M. (2022) Climate simulations: recognize
the ‘hot model’ problem, *Nature*, 605, https://www.nature.com/articles/d41586-022-01192-2 https://doi.org/10.1038/d41586-022-01192-2 PMID: 35508771

Hou, Y., Guo, H., Yang, Y., & Liu, W. (2023). Global evaluation of runoff simulation from climate, hydrological and land surface models. *Water Resources Research*, 59, e2021WR031817.

https://doi.org/10.1029/2021WR031817

Lehner, F. et al. (2019) The potential to reduce uncertainty in regional runoff projections from climate models, *Nature Climate Change*, 9, 926-933, https://doi.org/10.1038/s41558-019-0639-x

Lehner, F., et al. (2020) Partitioning climate projection uncertainty with multiple large ensembles and CMIP5/6. *Earth Syst. Dyn.*, 11, 491–508, https://doi.org/10.5194/esd-11-491-2020

Li, D., M. L. Wrzesien, M. Durand, J. Adam, and D. P. Lettenmaier (2017) How much runoff originates as snow in the western United States, and how will that change in the future?, *Geophys. Res. Lett.*, 44, 6163–6172, doi:10.1002/2017GL073551.

Qasmi, S., Ribes, A. (2022) Reducing uncertainty in local climate projections, *Research square*, 8, 41, doi :10.21203/rs.3.rs-364943/v1

Ribes, A., J. Boé, S. Qasmi, B. Dubuisson, H. Douville, and L. Terray (2022) An updated assessment of past and future warming over France based on a regional observational constraint, *Earth Syst. Dyn*., 13, 1397–1415, https://doi.org/10.5194/esd-13-1397-2022

Ribes, A., Qasmi, S., Gillett, N. (2021) Making climate projections conditional on historical observations. *Science Advances*, 7, eabc0671. https://doi.org/10.1126/sciadv.abc0671 PMID: 33523939

Sanderson, B. M., Pendergrass, A. G., Koven, C. D., Brient, F., Booth, B. B. B., Fisher, R. A., & Knutti, R. (2021) The potential for structural errors in emergent constraints, *Earth Syst. Dyn.*, 12, 899–918, https://doi.org/10.5194/esd-12-899-2021

Tang, Q. & Lettenmaier, D.P. (2012) 21st century runoff sensitivities of major global river basins. *Geophys. Res. Lett.*, 39, L06403, doi:10.1029/2011GL050834

Zhang, X., Tang, Q., Liu, X., Leng, G., & Di, C. (2018). Nonlinearity of runoff response to global mean temperature change over major global river basins. *Geophys. Res. Lett.*, 45, 6109–6116. https://doi.org/10.1029/2018GL078646

Revisiting the potential to narrow model uncertainty in the projections of Arctic runoff

Emma Dutot[1] and Hervé Douville[2*]

[1] Météo-France, Lyon, France

[2] Centre National de Recherches Météorologiques, Université de Toulouse, Météo-France, CNRS, Toulouse, France

* Corresponding author : herve.douville@meteo.fr

Abstract

Despite multiple advances in the understanding of the water cycle intensification in a warmer climate, climate models still diverge in their hydrological projections. Here we constrain annual runoff projections over individual and aggregated Arctic river basins. For this purpose, we use two ensembles of global climate models and two statistical methods: a regression scheme assuming similar runoff sensitivities at interannual versus climate change timescales, and a Bayesian method where models are used to derive a posterior runoff response conditional to historical observations. While both techniques are shown to narrow model uncertainties, more or less substantially depending on rivers, the Bayesian method is less sensitive to the choice of the model ensemble and is more skilful when tested with synthetic observations. It has also been applied over the whole Arctic watershed, showing so far a limited narrowing of the inter-model spread, but its skill will further improve with increasing climate change.

Plain language summary:

Despite considerable progress in understanding the intensification of the water cycle in response to global warming, projections of changes in the various components of the continental water balance remain highly model-dependent. This is particularly the case for the evolution of runoff, which remains very sensitive to the way atmospheric and continental processes are represented. Here we evaluate and compare two statistical methods to constrain climate projections of annual runoff at high latitudes via available or synthetic observations. Our Bayesian method, based on historical observations, is found to be more robust than a previous method based on observed interannual climate variability. It underlines the importance of having reliable flow reconstructions to constrain runoff projections and a careful use of more empirical emergent constraints which can lead to overconfident projections.

Key points :

- Runoff sensitivity to temperature depends on the considered timescale.
- Observational constraints based on interannual variability can thus lead to overconfident projections.
- Observational constraints based on historical trends can be effective if such trends are driven by human emissions of greenhouse gasses.

1. Introduction

According to the latest IPCC report, hydrological projections at the regional scale are still very uncertain despite a robust theoretical understanding of the water cycle intensification at the global scale (AR6 WG1, Douville et al., 2021). More specifically, there is low confidence in the sign and magnitude of projected changes in global land runoff in all illustrative Shared Socio-economic Pathway (SSP) scenarios scrutinized in the sixth phase of the Coupled Model Intercomparison Project (CMIP6). There are however some regional exceptions, especially in the northern high latitudes where there is high confidence that projected increases in precipitation amount and intensity will be associated with increased runoff throughout the 21st century.

Much of the uncertainty in hydrological projections arises from global climate model (GCM) uncertainty rather than emission scenario uncertainty, while internal climate variability also plays a key role in the next two decades (Lehner et al., 2020 ; Douville et al., 2021). Moreover, CMIP6 often shows larger model uncertainties than CMIP5 (Lehner et al., 2020), which may be partly due to a wider range of climate sensitivity (Hausfather et al., 2022). Yet, this explanation may not hold for runoff whose regional response does not scale with global warming across multiple models given the influence of other potential drivers, including model-dependent changes in large-scale circulation (Douville et al., 2022; Elbaum et al., 2022).

A recent global evaluation of runoff simulations, from both CMIP6 GCMs and off-line global hydrological models, suggests that models generally perform better in non-cold environments and that the multi-model ensemble average is not systematically an effective way to reduce bias compared to individual models (Hou et al., 2023). These results further support previous findings based on CMIP5 GCMs that urged caution in the direct use of climate model runoff for hydrological and water security applications (Lehner et al., 2019). Regional climate models at higher spatial resolution are extremely useful to project runoff in small watersheds but do not necessarily provide more reliable projections in large-scale river basins since they are strongly constrained by their driving GCMs. Non-linearities in the runoff response to global warming may also challenge the use of simple pattern-scaling techniques to produce reliable hydrological projections at the

regional scale (Zhang et al., 2018 ; Douville et al., 2021).

In response to these difficulties, an increasingly popular approach is to use emergent constraints to link future changes to present-day or recent climate features across multiple GCMs. Unfortunately, most emergent constraints that have been proposed to narrow uncertainties in CMIP5 projections generally do not perform as well for CMIP6 models (Sanderson et al., 2021). These results support previous studies concluding that emergent constraints should be based on an independently verifiable physical mechanism (e.g., Hall et al., 2006) and should be considered with great caution when there is no such dominant single mechanism but, as often, a plurality of poorly-constrained mechanisms due to possible gaps in both models and observations. Emergent constraints have thus the potential to produce overconfident projections, especially when processes are represented in a common way throughout the GCM ensemble (Sanderson et al., 2021). Moreover, they sometimes assume a time-invariance of climate feedbacks so that the mechanisms that dominate future climate changes are considered to be similar and observable on shorter timescales such as the annual cycle (Hall et al., 2006) or interannual variability (Douville et al., 2006 ; Lehner et al., 2019). This is a strong hypothesis given the growing evidence that climate feedbacks may not only vary across timescales (e.g., Dessler and Forster, 2018) but may also be time-dependent on climate change timescales (e.g., Andrews et al., 2022).

Only few studies have quantified regional runoff projections in CMIP models, generally demonstrating a wide inter-model spread but a strong consistency in the ensemble mean model behavior across emissions scenarios or model generations (e.g., Tang et al., 2012 ; Zhang et al., 2018 ; Giuntoli et al., 2018). Even less studies have tried to go beyond the "one model one vote" approach in order to assess more carefully the regional runoff sensitivity (e.g., Lehner et al., 2019 ; hereafter L19). Unlike in off-line global hydrological models, runoff sensitivity is rarely explicitly tuned in GCMs, thus offering an opportunity to develop an observational constraint unaffected by model calibration efforts. In L19, there are however two key sensitivities for which the time-invariance assumption has been made: the precipitation sensitivity of runoff (sometimes termed ‘runoff elasticity’ ; e.g., Tang and Lettenmaier, 2012), defined as the percent change in runoff induced by a 1% change in precipitation ($\Delta Q/\Delta P$); and the temperature sensitivity of runoff, defined as the percent change in runoff for a 1°C change in temperature ($\Delta Q/\Delta T$). Such metrics can be diagnosed

from observations but their relevance is not guaranteed when it comes to constrain runoff projections across the whole 21st century.

In the present study, we illustrate this issue by revisiting L19 and downrating the potential to constrain the runoff response to climate change from the observed interannual variability. As in L19, the focus is on the northern high latitudes (although here including both North America and northern Eurasia) where a significant fraction of runoff originates as snowmelt and may be thus particularly sensitive to global warming (e.g., Li et al, 2017). After a brief description of the various datasets and statistical methods (Section 2), the limitations of the L19 method are first shown by using two definitions of runoff and two generations of GCMs in a high-emission scenario over four northern high-latitude river basins (Section 3). Moreover, the results of the L19 method are compared to those obtained with an alternative statistical technique recently developed at CNRM, the so-called Krigging for Climate Change technique (hereafter KCC; Ribes et al., 2020 ; Qasmi and Ribes, 2022). For the sake of a fair comparison, a pseudo-observation setting is also used in which one GCM is discarded from the CMIP ensemble and used to provide both historical and future data. Finally (Section 4), the reasons for the better behavior of the KCC method are briefly discussed and the method is also applied to constrain runoff projections over the aggregated "Arctic" basin (i.e., river basins whose outlet is in the Arctic ocean or neighboring seas).

2. Data and methods

We work with three main variables, total precipitation (P), near-surface air temperature (T) and total runoff (R), that have been averaged over large river basin and throughout the hydrological year (from October of the year N to September of the year N+1, as in L19). The focus is on the northern high-latitudes where there is an amplification of global warming, a human-induced increase in annual mean precipitation (Douville et al., 2021) and an expected climate change signal in the observed runoff reconstructions.

2.1 Observed and simulated data

The main observed monthly datasets used are precipitation from GPCC (1901-2018), near-surface air temperature from CRU_TS4 (1901-2019) or HadCRUT5 (1850-2020), and runoff from GRUN (1902-2013). Note that GRUN does not provide "real" observations, but is based on a machine learning method that estimates runoff from T and P observations (Ghiggi et al., 2019). The model is trained with monthly observations at the scale of relatively small catchment areas and then is applied on gridded data from reanalysis.

For simulated data, historical and future simulations (1850-2100) are taken from the last two phases of the Coupled Model Intercomparison Project (CMIP5 and CMIP6). We worked with the comparable although slightly different high-emission scenarios (RCP8.5 and SSP5-8.5 for CMIP5 and CMIP6 respectively) in order to maximize the signal to noise ratio. We distinguished several multi-model ensembles: ensembles with only one realization (data from the first available member of each model, we named these ensembles rcp81 and ssp51 respectively); a CMIP6 ensemble with at least 3 members but less models although we only used the average of all available realizations for each model (named ssp50), and a few mono-model ensembles (ssp5x) with 25 individual members.

2.2 Statistical methods used to constrain basin-scale runoff projections

The CMIP projections of 12-month runoff averaged from October to next September (the so-called water year) have been constrained with two different statistical methods. The first method is the L19 method depicted in Lehner et al. (2019) and based on runoff predictions from a multiple linear regression whose coefficients are constrained with the observed interannual variability. In the end, we can compute the basin-wide average runoff anomaly $\Delta R = a_0*\Delta P + b_0*\Delta T$, where a_0 and b_0 are the regression coefficients computed from observations (observed interannual variability) over the common 1902-2013 period. This period will be also referred to as the training period when L19 will be evaluated with model outputs as synthetic observations.

The second method (KCC) was developed by Ribes et al. (2021) and Qasmi and Ribes (2022) and is based

on Bayesian statistics where a *prior* distribution, $\pi(x)$, of the forced response to anthropogenic forcings is first derived from raw model outputs and then constrained directly with observations of one or more variables (here both observed global mean surface temperature and reconstructed basin-wide average runoff). The *prior* is estimated using a Generalized Additive Model (GAM, assuming the additivity of the model responses to individual forcings) and a simple Energy Budget Model (EBM, allowing us to diagnose the runoff response to volcanic eruptions; for more details, see supplementary materials from Ribes et al., 2021). For the sake of simplicity, this *prior* is assumed to follow a normal distribution and thus only needs an estimate of the ensemble mean and spread. Next, observations y are used to derive a *posterior* distribution (after constraint). We suppose that observations can be described as: $y=\mathcal{H}*x+\varepsilon$, where $\mathcal{H}$ is a pseudo-observation operator allowing to extract the part of x observed in y and ε corresponds to internal variability and observational errors (if available). Since $\pi(x)$ and ε are assumed to follow normal distributions, the *posterior* can be easily derived using the Gaussian conditioning theorem (for more information, see supplementary materials from Ribes et al., 2021).

These two statistical methods are thus very different. L19 is using the observed internal variability to constrain the regression coefficient of a multiple regression scheme aimed at predicting the long-term evolution of runoff from the evolution of the basin-wide average precipitation and near-surface temperature. In contrast, KCC is using directly the observed evolution of runoff and/or of global warming over the historical period to constrain the forced response of the simulated runoff with a Bayesian statistical method. Given the use of interannual runoff fluctuations to fit the regression scheme, L19 needs individual realizations of each model. For the sake of simplicity and of a fair comparison between the two methods, we thus start to work with the ssp51 and rcp81 multi-model ensembles. Yet, KCC focuses on the forced response and is thus expected to provide even better results when using as many realizations as possible for each model, hence also the use of the ssp50 ensemble based on the average of all members for each model.

2.3. Pseudo-observations and related scores

To assess and compare the reliability of the L19 and KCC methods, we decided to work with

pseudo-observations (as done in Ribes et al., 2021). The idea is that we extract one of the models and we use it as synthetic observations of both past and future variations of T, P and R. In the end, after constraining our remaining projections using the early pseudo-record (1901-2013), we can check if they are getting closer to our R pseudo-observations over the end of the 21st century compared to the raw (i.e., unconstrained) runoff projections. Each model provides successively the reference pseudo-observations, so that we can compute several probabilistic scores to assess the quality and reliability of our constrained projections. Beyond the inter-model spread or the multi-model 90% confidence interval, two main scores were computed in this study : the coverage probability (CP) and the CRPS change.

- Coverage probability (CP): for each iteration, we check that the future pseudo-observations for the 2081-2100 period (water-year mean anomalies over this period) are in the 5-95% confidence interval of the constrained projections. In the end, we get a CP percentage corresponding to the number of reference models (POBS) in this confidence interval, which should be as close to 90% as possible.
- CRPS is a probabilistic score quantifying the distance between an observation (here corresponding to "pseudo-observed" 20-water-year mean runoff anomalies) and the distribution of predicted anomalies over the same 2081-2100 period. The closer the observation is to the predicted distribution, the closer the CRPS is to 0. We compared the CRPS computed with the model distribution before versus after constraint. Then we computed the relative change in CRPS: a negative change will thus mean that observation and predictions are closer after constraint (i.e., improved distribution). For each basin, we averaged this CRPS change after using successively each model as pseudo-observations.

3. Results

Our main purpose here is first to constrain the ssp51 and rcp81 ensembles of runoff projections with observations, using two different statistical methods. We will compare the robustness and reliability of the two methods by first checking the sensitivity to the chosen model ensemble (CMIP6 vs CMIP5), and then computing statistical scores using the idealized pseudo-observation framework.

3.1 Assessing the robustness of the L19 method

In Lehner et al., 2019, the L19 method was applied to CMIP5 models and runoff was estimated as P-E (precipitation - evapotranspiration), thus assuming no change in water storage. In the present study, we work with CMIP6 models and genuine simulated runoff outputs. In agreement with L19 (cf. their Fig.1), Fig.S1 in the supplementary information (SI) shows that runoff anomalies computed with a multiple linear regression (using both P and T predictors) performs better than a simple regression (using P only) to predict runoff anomalies over the training period (1901-2013). Note that other regressions could be also tested but would not be as effective or based on less reliable predictors in the instrumental record. We thus choose to work with the same basin-wide average P and T predictors as in L19. A similar figure but using P-E instead of R (as in Lehner et al., 2019) is shown in SI (Fig.S2). Not surprisingly, regressions show an improved skill when using P-E rather than R. Thus, using P-E as an approximation of R can lead to overconfident projections when using the L19 method. Thereafter we will work with the R variable averaged throughout the water year, for which river the observed discharge provides reasonable estimates.

The L19 method apparently succeeds in narrowing uncertainty in runoff projections. Indeed, the 5-95% confidence interval is reduced after the L19 constraint (black and green line, Fig.1) compared to the simulated and raw predicted anomalies at the end of the 21st century. Overall, even when uncertainties in the observed regression coefficients are accounted for (green thin line), we obtain a systematic reduction of the 90% confidence interval. Similar qualitative results are obtained when using P-E as a surrogate of runoff (Fig.S3). Yet, our results are basin-dependent and also sensitive to the choice of the multi-model ensemble as shown by the comparison between CMIP6 (Fig.1) and CMIP5 (Fig.S4).

To sum up, the L19 method seems to succeed in narrowing the uncertainties in runoff projections over the four selected northern high-latitude river basins : Columbia, Kolyma, Lena and Mackenzie. However, these results rely on severe assumptions that have not been tested properly and may lead to overconfident projections. This is the reason why we will further evaluate the reliability of the L19 method by using synthetic observations of both past and future climate in Section 3.3.

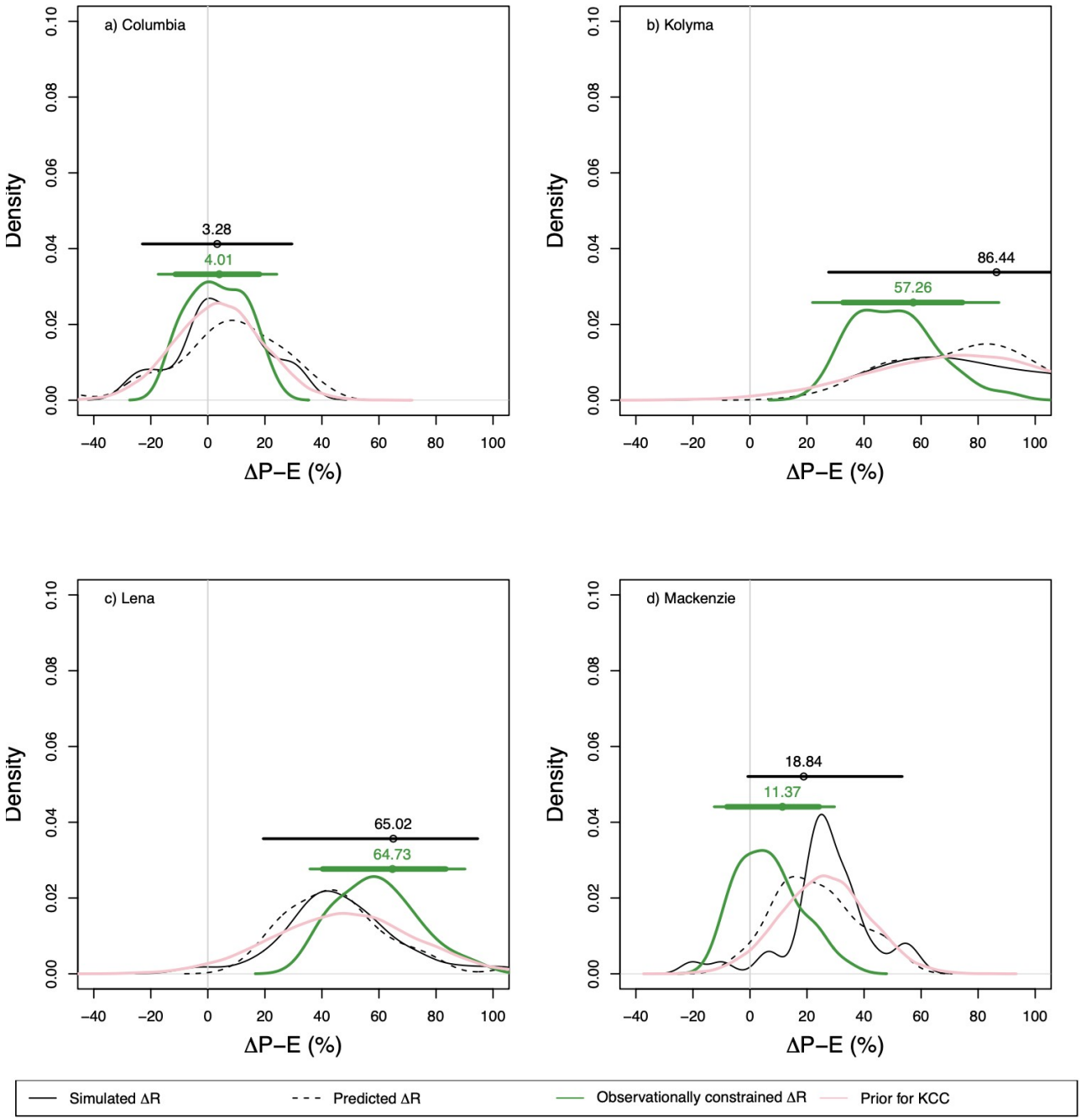

Figure 1: Constrained versus unconstrained distributions of water-year mean relative runoff anomalies (%) from the CMIP6 model ensemble under the SSP5-8.5 high-emission scenario : a) Columbia, b) Kolyma, c) Lena, and d) Mackenzie. Kernel density functions of runoff anomalies as simulated by CMIP6 models and as predicted and constrained by the L19 method (runoff anomalies are averaged over the 2081-2100 period). The KCC prior distribution is also shown ($\pi(x)$ in red) in order to highlight that the normal distribution assumption has a significant impact on the prior distribution given the limited size of the CMIP6 ensemble. Predicted runoff anomalies are those computed from the L19 multiple regression but without constraining the regression coefficients with observations. The black and green horizontal lines

represent the 5-95% confidence interval of the simulated and the L19 constrained projections respectively. The green confidence interval includes uncertainties in the regression coefficients as estimated with a bootstrap method (green thin line).

3.2 Assessing the robustness of the KCC method

The KCC method was also applied to constrain both CMIP6 and CMIP5 annual runoff projections. Two metrics were used as observational constraints: the global mean near-surface air temperature from HadCRUT5 (including an estimate of observational errors) and/or the basin-wide average runoff from GRUN (without error estimates and based on reconstructed runoff from observations rather than genuine river discharge measurements). For the four selected river basins, the CMIP6 inter-model spread of the distribution is reduced by around 10 to 35% in 2100 after applying the KCC observational constraint (Fig.2). Similar results are obtained with the CMIP5 models (Fig.S5). The limited narrowing of model uncertainty, compared to the use of L19 and in particular for the Columbia river also considered in L19, is consistent with the lack of an obvious long-term trend in the GRUN reconstruction. When comparing the constrained projections between ssp51 (Fig.2) and ssp50 (with multiple members for each model, Fig.S6), similar results are found thereby suggesting that internal climate variability is not responsible for the apparently lower performance of KCC.

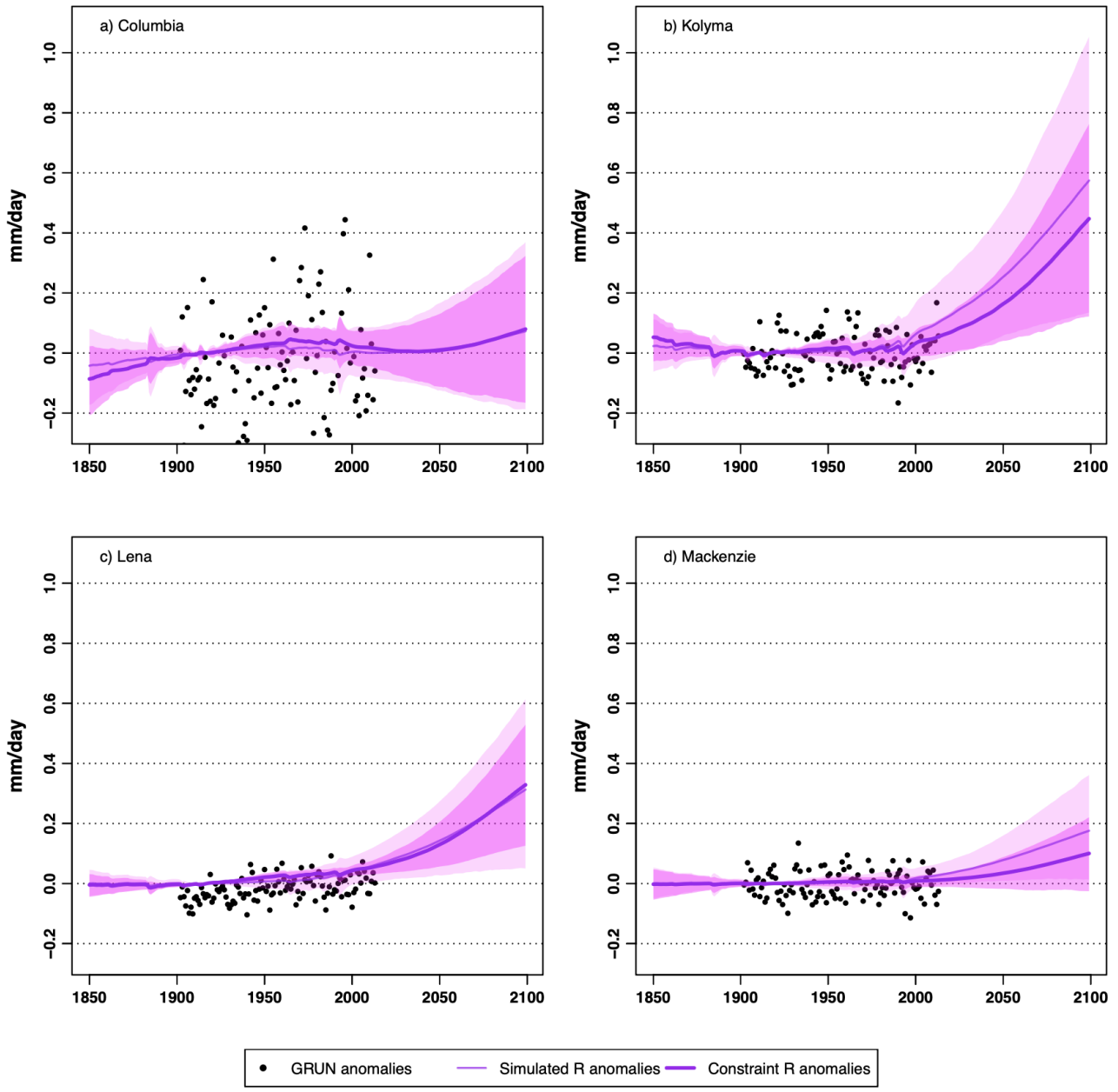

Figure 2: Constrained versus unconstrained water-year runoff anomalies (mm/day) using the ssp51 ensemble (a single realization of each CMIP6 model under the SSP5-8.5 scenario). Black dots correspond to the GRUN water-year runoff anomalies. Both GRUN runoff and HadCRUT5 GSAT observations are used to constrain the simulated runoff anomalies. The thick lines denote the best estimate of each distribution (i.e., the ensemble mean) while shadings denote the corresponding 5-95% confidence intervals.

The KCC method was also applied using only either the global mean surface temperature observations (GSAT from the HadCRUT5 dataset, consisting of 200 members to account for observational uncertainties) or the basin-wide runoff GRUN reconstruction (assuming no observational uncertainty). When using only

GSAT observations, the inter-model spread is not much reduced, thereby supporting the claim that regional hydrological changes are not heavily constrained by global warming (Douville et al., 2022). This finding highlights the need for monitoring river discharge and providing estimates of water withdrawals versus natural runoff, including observational uncertainties, for better constraining runoff projections.

To sum up, KCC was also applied to Columbia, Kolyma, Lena and Mackenzie and led to a weaker (by 10% to 35%) but more robust (CMIP5 versus CMIP6) narrowing of model uncertainty in the highest emissions scenarios and using a single member for each model. Yet, the performance of KCC is expected to improve with the emerging influence of climate change in the observed time series, in contrast to L19 where longer observed time series will only have a limited impact on the observed regression coefficients dominated by interannual variability. Moreover, a fair comparison between KCC and L19 necessitates a different setting where physically-consistent pseudo-observations are derived from a randomly-chosen climate model in order to assess the constrained projections based on the other models.

3.3 Validation of these results with pseudo-observations

We showed that both L19 and KCC methods have the potential to reduce model uncertainty on projected runoff anomalies in a warmer climate. Yet, we do not know so far the extent to which these results are reliable. As in Ribes et al.(2021), we now use a pseudo-observation framework and compute three probabilistic scores (CP, CRPS change and reduction of spread) in order to assess the robustness and reliability of our methods (see section 2.2 for details). Results are compared over the four selected watersheds in order to get an overall picture of the relative performance of the L19 and KCC methods (table S1, S2 and S3). As a reminder, and beyond the expected reduction in the inter-model spread, CP has to be as near to 90% as possible and CRPS has to be as low (close to zero) as possible after constraining the projections. Therefore, the change in CRPS and spread have to be negative for the constraint to be reliable and effective.

Although leading to a fairly similar systematic reduction of uncertainty, the L19 and KCC methods show

contrasted scores in terms of both CP and CRPS. For the L19 method, CP is far from 90% and changes in CRPS are systematically positive, thus providing clear evidence that the method leads to overconfident projections, likely due to a timescale-dependent temperature effect on runoff (cf. Fig.S7 and S8). For the KCC method, CP is above 80% for three out of four basins (table S3) and CRPS changes are either negative or slightly positive. While these results suggest a cautious use of both methods, they clearly emphasize the superiority of the Bayesian KCC technique compared to the more empirical L19 emergent constraint.

Regarding the KCC method, scores are generally (and as expected) slightly improved if one uses several rather than just one realization to assess the forced response of each model (ssp50, compared to ssp51; better spread reduction, overall better change in CRPS but slightly worse CP). Yet, they are still basin-dependent and better when using the local (GRUN) constraint only (not shown here). This finding highlights again that the observed global historical warming has little positive influence on the constrained runoff and further supports the cautious statement of Douville et al. (2022) regarding the scalability of hydrological changes with global warming across multiple models.

Note that we also tried to constrain the basin-scale projections of precipitation and temperature using KCC, before applying the L19 method (multiple linear regression with observationally-constrained regression coefficients) to constrain the projections of runoff. However, this combined method led to even more overconfidence than applying the L19 method only when tested with pseudo-observations, so that the results were not included in this article.

To sum up, scores are overall better when using the KCC method to constrain runoff projections with pseudo-observations compared to the use of the more empirical L19 method. The overconfidence of L19-constrained runoff projections arises primarily from the contrasted runoff sensitivity to temperature and precipitation at climate change versus interannual timescale (cf. Fig.S7 and S8).

4. Discussion

Two statistical methods were applied with available observations, and compared in a more idealized setting, to reduce the uncertainty about projected northern-latitude runoff changes throughout the 21st century at the basin-scale the L19 method (Lehner et al., 2019) and the KCC method (Ribes et al., 2021). After using pseudo-observations to compute statistical scores and assess the reliability of our results, we found that the L19 method is systematically overconfident, which means that there is a greater than 10% probability for the actual runoff evolution to lie outside the constrained 5-95% confidence interval. Generally speaking, the KCC method leads to more reliability and better scores, though variable from one basin to another depending on the observed trend in the GRUN natural runoff reconstructions (not accounting for a direct human influence through water withdrawals and river management). The overconfidence of the L19 results is primarily due to a wrong hypothesis regarding the runoff sensitivity to temperature and precipitation at different timescales (e.g., Zhang et al., 2022). Our results thus illustrate the danger of using emergent constraints without having first tested them in an idealized context (Sanderson et al., 2021). Furthermore, they highlight the potential of more direct observational constraints whose power should increase with further climate change and stronger signal to noise ratio in observed runoff time series.

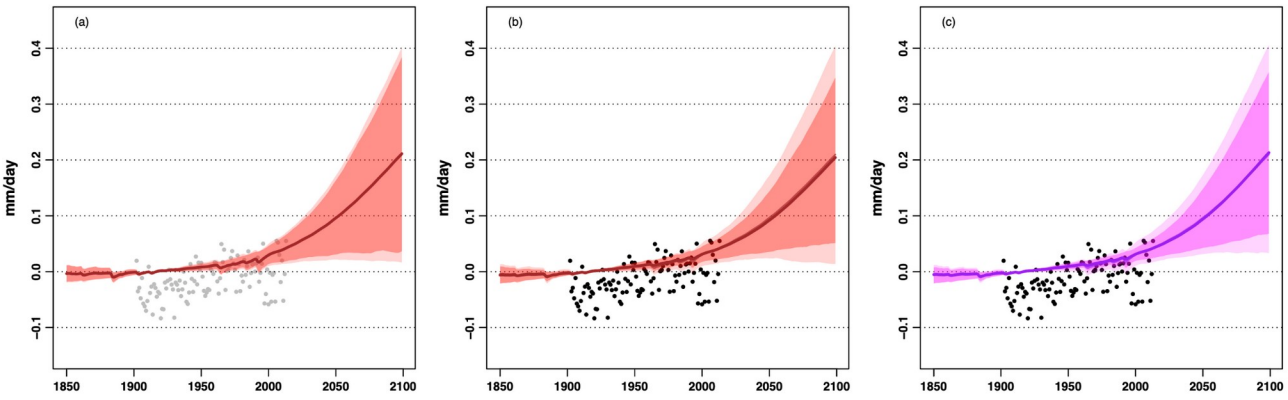

Figure 3 : Constrained versus unconstrained aggregated Arctic water-year runoff anomalies (mm/day) using the ssp50 ensemble a) KCC constraint using only the HadCRUT5 GSAT observations. **b**) KCC constraint using only the GRUN runoff reconstruction. **c**) KCC constraint using both HadCRUT5 and GRUN observational constraints. Black dots correspond to the GRUN water-year anomalies; they are colored in grey (rather than black) when GRUN is not used to constrain the projections. The thick lines (here

superimposed) denote the best estimate of each distribution (i.e., the ensemble mean) while shadings denote the corresponding 5-95% confidence intervals.

To conclude this study, we have thus decided to apply the KCC method to the aggregated "Arctic" basin by merging all watersheds whose outlets are in the Arctic Ocean and nearby seas. Results of the individual and combined constraints are shown Fig.3. Overall, the constraint using both GSAT and runoff observations (GRUN and HadCRUT5) allows KCC to reduce the CMIP6 inter-model spread in Arctic runoff by 22% (Fig.3.c) in 2100 (even more in the early 21st century). As expected when increasing the signal to noise ratio through spatial aggregation, this combined constraint is more efficient than the individual ones using either HadCRUT5 or GRUN reconstructions. Scores with pseudo-observations for the whole Arctic basin (not shown) were also computed. They are quite promising and suggest even better results after one more decade of observations. Our study thus emphasizes the need of more reliable and routinely updated runoff observations (compared to GRUN for instance) to constrain model projections that, unfortunately, do not show a spread reduction from one model generation to the next despite the sustained efforts of global modeling centers to improve and evaluate such models.

Data Availability Statement

We did not use new data in the present study. The CMIP5 and CMIP6 monthly mean model outputs are available on the ESGF archive at https://esgf-node.llnl.gov/, GRUN reconstruction of monthly runoff are available at https://www.researchcollection.ethz.ch/handle/20.500.11850/324386, HadCRUT5 global mean surface temperature is available at https://www.metoffice.gov.uk/hadobs/hadcrut5/.

Code Availability Statement

The KCC statistical package for observational constraint is available on gitlab at https://gitlab.com/saidqasmi/KCC. Other codes for CMIP5 and CMIP6 data curation and visualization are based on the CliMAF package available at: https://climaf.readthedocs.io/en.

References :

Andrews, T., Bodas-Salcedo, A., Gregory, J. M., Dong, Y., Armour, K. C., Paynter, D., et al. (2022). On the effect of historical SST patterns on radiative feedback. Journal of Geophysical Research: Atmospheres, 127, e2022JD036675. https://doi.org/10.1029/2022JD036675

Dessler, A. E., & Forster, P. M. (2018). An estimate of equilibrium climate sensitivity from interannual variability. *Journal of Geophysical Research:Atmospheres*,123, 8634–8645. https://doi.org/10.1029/2018JD02848

Douville, H., Salas-y-Mélia, D., Tyteca, S. (2006) On the tropical origin of uncertainties in the global land precipitation response to global warming. *Clim. Dyn.,* 26, 367–385, doi :10.1007/s00382-005-0088-2

Douville, H., Raghavan, K., Renwick, J., Allan, R.P., Arias, P.A., Barlow, M. et al. (2021) Water Cycle Changes. In Climate Change 2021: The Physical Science Basis. *Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 1055–1210, https://doi.org/10.1017/9781009157896.010

Douville, H., Allan, R.P., Arias, P.A., Betts, R.A., Caretta, M.A., Cherchi, A., et al. (2022) Water remains a blind spot in climate change policies. *PLOS Water*, 1(12): e0000058. https://doi.org/10.1371/journal.pwat.0000058

Douville, H., Qasmi, S., Ribes, A., Bock, O. (2022) Global warming at near-constant relative humidity further supported by recent in situ observations. *Communications Earth & Environment*, 3, 237, https://doi.org/10.1038/s43247-022-00561-z

Elbaum, E., Garfinkel, C.I., Adam, O., Morin, E., Rostkier-Edelstein, D., Dayan, U. (2022) Uncertainty in projected changes in precipitation minus evaporation: dominant role of dynamic circulation changes and weak role for thermodynamic changes. *Geophys. Res. Lett.,* https://doi.org/10.1029/2022GL097725

Ghiggi, G., Humphrey, V., Seneviratne, S. I. and Gudmundsson, L. (2019) GRUN : an observation-based global gridded runoff dataset from 1902 to 2014. *Earth System Science Data*, 11, 1655–1674, https://doi.org/10.5194/essd-11-1655-2019

Giuntoli, I., Villarini, G., Prudhomme, C., Hannah, D.M. (2018) Uncertainties in projected runoff over the conterminous
United States. *Climatic Change*, 150, 149–162, https://doi.org/10.1007/s10584-018-2280-5

Hall, A., and X. Qu (2006), Using the present-day seasonal cycle to constrain climate sensitivity: A case study of snow albedo feedback, *Geophys. Res. Lett.*, 33, 1550–1568, doi:10.1029/2005GL025127

Hausfather, Z., Marvel, K., Schmidt, GA., Nielsen-Gammon, JW., Zelinka, M. (2022) Climate simulations: recognize
the ‘hot model’ problem, *Nature*, 605, https://www.nature.com/articles/d41586-022-01192-2 https://doi.org/10.1038/d41586-022-01192-2 PMID: 35508771

Hou, Y., Guo, H., Yang, Y., & Liu, W. (2023). Global evaluation of runoff simulation from climate, hydrological and land surface models. *Water Resources Research*, 59, e2021WR031817.

https://doi.org/10.1029/2021WR031817

Lehner, F. et al. (2019) The potential to reduce uncertainty in regional runoff projections from climate models, *Nature Climate Change*, 9, 926-933, https://doi.org/10.1038/s41558-019-0639-x

Lehner, F., et al. (2020) Partitioning climate projection uncertainty with multiple large ensembles and CMIP5/6. *Earth Syst. Dyn.*, 11, 491–508, https://doi.org/10.5194/esd-11-491-2020

Li, D., M. L. Wrzesien, M. Durand, J. Adam, and D. P. Lettenmaier (2017) How much runoff originates as snow in the western United States, and how will that change in the future?, *Geophys. Res. Lett.*, 44, 6163–6172, doi:10.1002/2017GL073551.

Qasmi, S., Ribes, A. (2022) Reducing uncertainty in local climate projections, *Research square*, 8, 41, doi :10.21203/rs.3.rs-364943/v1

Ribes, A., J. Boé, S. Qasmi, B. Dubuisson, H. Douville, and L. Terray (2022) An updated assessment of past and future warming over France based on a regional observational constraint, *Earth Syst. Dyn*., 13, 1397–1415, https://doi.org/10.5194/esd-13-1397-2022

Ribes, A., Qasmi, S., Gillett, N. (2021) Making climate projections conditional on historical observations. *Science Advances*, 7, eabc0671. https://doi.org/10.1126/sciadv.abc0671 PMID: 33523939

Sanderson, B. M., Pendergrass, A. G., Koven, C. D., Brient, F., Booth, B. B. B., Fisher, R. A., & Knutti, R. (2021) The potential for structural errors in emergent constraints, *Earth Syst. Dyn.*, 12, 899–918, https://doi.org/10.5194/esd-12-899-2021

Tang, Q. & Lettenmaier, D.P. (2012) 21st century runoff sensitivities of major global river basins. *Geophys. Res. Lett.*, 39, L06403, doi:10.1029/2011GL050834

Zhang, X., Tang, Q., Liu, X., Leng, G., & Di, C. (2018). Nonlinearity of runoff response to global mean temperature change over major global river basins. *Geophys. Res. Lett.*, 45, 6109–6116. https://doi.org/10.1029/2018GL078646

Geophysical Research Letters

Supporting Information for

Revisiting the potential to narrow uncertainty in the projections of Arctic runoff

E. Dutot[1], H. Douville[1]

[1] Centre National de Recherches Météorologiques, Université de Toulouse, Météo-France, CNRS, 42 Avenue Gaspard Coriolis, 31057 Toulouse, France

Content of this file:

- Table S1 and S2 compare the performance of the L19 and KCC statistical methods at the basin scale using pseudo-observations derived from a single realization of one out of the CMIP6 models.

- Table S2 and S3 compare the performance of the KCC statistical method at the basin scale using pseudo-observations derived from single versus multiple realizations of one out of the CMIP6 models.

- Fig. S1 and S2 compare the skill of the L19 regression method to predict simulated anomalies depending on the considered variable (P-E instead of runoff).

- Fig. S3 and S4 (to be compared with Fig. 1) evaluate the sensitivity of the L19 method to the choice of the constrained variable (P-E instead of runoff) and of the model ensemble (CMIP5 instead of CMIP6).

- Fig. S5 and S6 (to be compared with Fig. 2) evaluate the sensitivity of the KCC method to the choice of the model ensemble (P-E instead of runoff) and of the model ensemble (CMIP5 instead of CMIP6).

- Fig. S7 and S8 compare the runoff sensitivity to temperature and precipitation at short versus long timescales and explain why the L19 results should be considered with caution.

Basin	Coverage probability (%)	Change in CRPS (%)	Spread reduction (%)
Columbia	69.7	9.06	-37.5
Kolyma	57.58	22.8	-36.74
Lena	69.7	7.74	-12.83
Mackenzie	78.79	58.06	-14.24

Table S1: Probabilistic scores of the L19 method using the ssp51 ensemble (a single realization of each CMIP6 model under the SSP5-8.5 high-emission scenario). For each river basin, CRPS change and spread reduction are averaged after using successively each CMIP6 model as pseudo-observations.

Basin	Coverage probability (%)	Change in CRPS (%)	Spread reduction (%)
Columbia	84.85	16.08	-13.44
Kolyma	87.88	0.17	-28.61
Lena	93.94	-6.25	-23.88
Mackenzie	84.85	-22.65	-23.29

Table S2: Probabilistic scores of the KCC method using the ssp51 ensemble (a single realization of each CMIP6 model under the SSP5-8.5 high-emission scenario) and two observational constraints. For each river basin, CRPS change and spread reduction are averaged after using successively each CMIP6 model as pseudo-observations.

Basin	Coverage probability (%)	Change in CRPS (%)	Spread reduction (%)
Columbia	82.76	4.86	-25.48
Kolyma	82.76	-18.75	-39.89
Lena	86.21	6.54	-30.64
Mackenzie	72.41	-10.31	-35.39

Table S3: Probabilistic scores of the KCC method using the ssp50 ensemble (multiple realizations of each CMIP6 model under the SSP5-8.5 high-emission scenario) and two observational constraints. For each river basin, CRPS change and spread reduction are averaged after using successively each CMIP6 model as pseudo-observations.

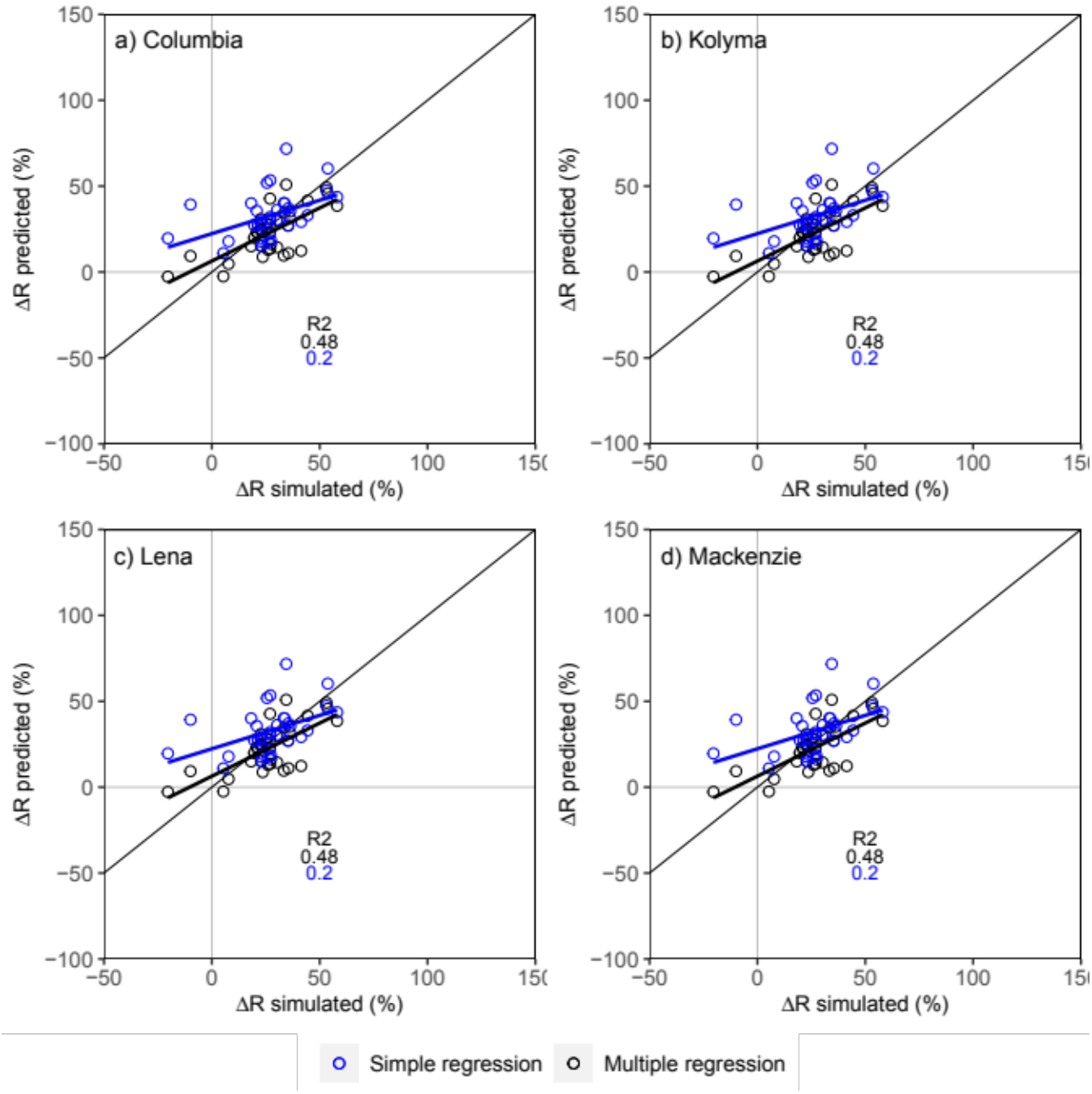

Figure S1: Scatterplots of predicted (L19) versus simulated (GCM) relative anomalies (%) of basin-scale water-year runoff in individual CMIP6 models under the SSP5-8.5 high-emission scenario: a) Columbia, b) Kolyma, c) Lena, d) Mackenzie. L19 runoff anomalies are computed from a simple (ΔR~ΔP) or multiple (ΔR~ΔP+ΔT) linear regression and plotted against the corresponding simulated anomalies. All anomalies are averaged over 2081-2100 relatively to the 1902-1930 baseline period. In each panel, R2 denotes coefficient of determination of the linear regression. The closer the regression line is from y=x (thin black solid line), the better the linear regression is. Not surprisingly, the multiple regression (black circles) is better than the simple regression (blues circles) at predicting the simulated anomalies.

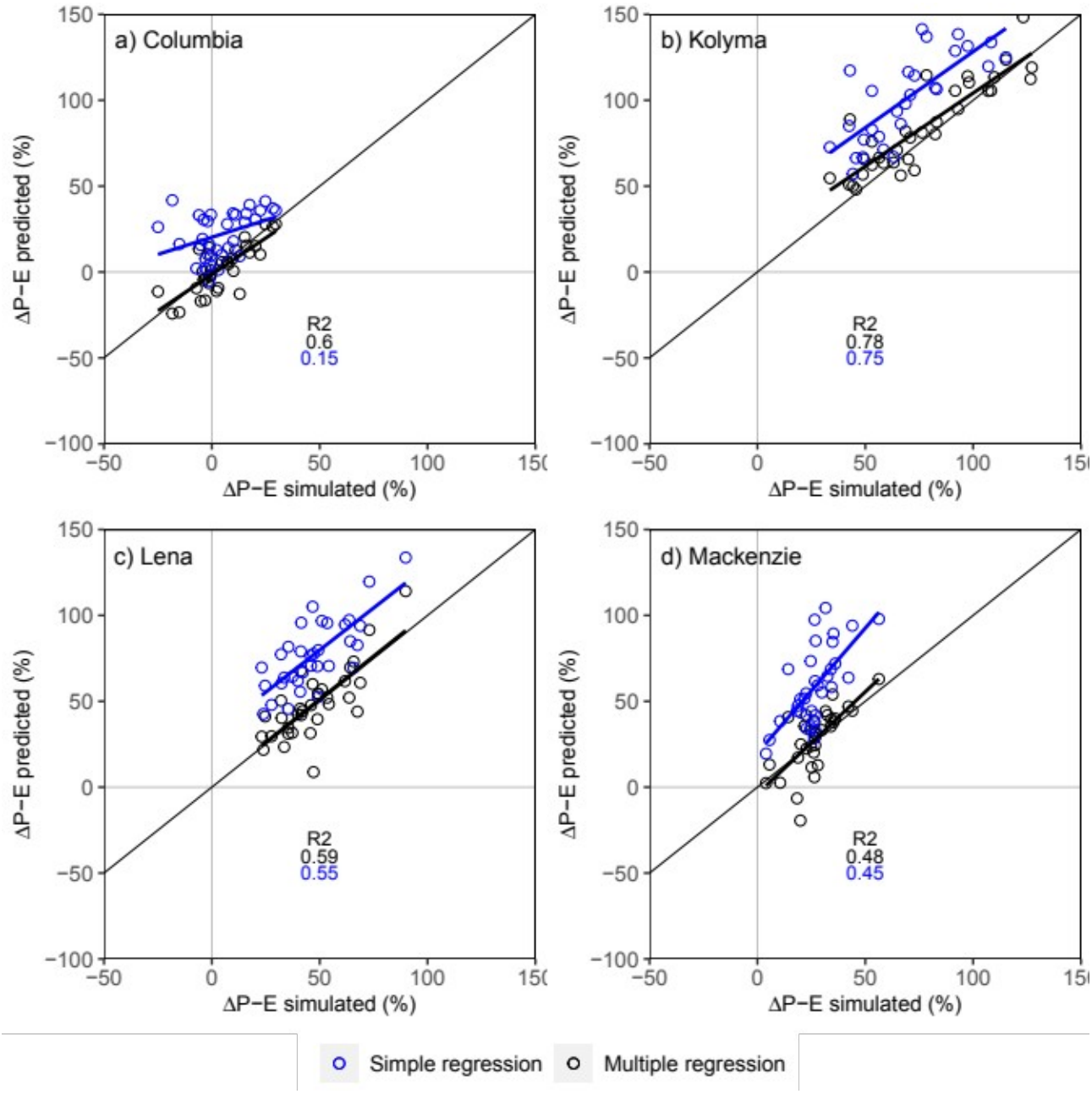

Figure S2: Same as Fig. S1 but using P-E (precipitation minus evapotranspiration) rather than runoff relative anomalies (%). Not surprisingly, it is easier to predict P-E rather than R anomalies with both simple (ΔP-ΔE~ΔP) and multiple (ΔP-ΔE~ΔP+ΔT) regressions given the strongly model-dependent simulated soil moisture anomalies in climate models. When temperature is not accounted for, the P-E simulated anomalies are systematically overestimated by the regression, thus highlighting the strong influence of global warming on surface evapotranspiration. Using P-E as a surrogate for water-year runoff therefore leads to overconfident projections.

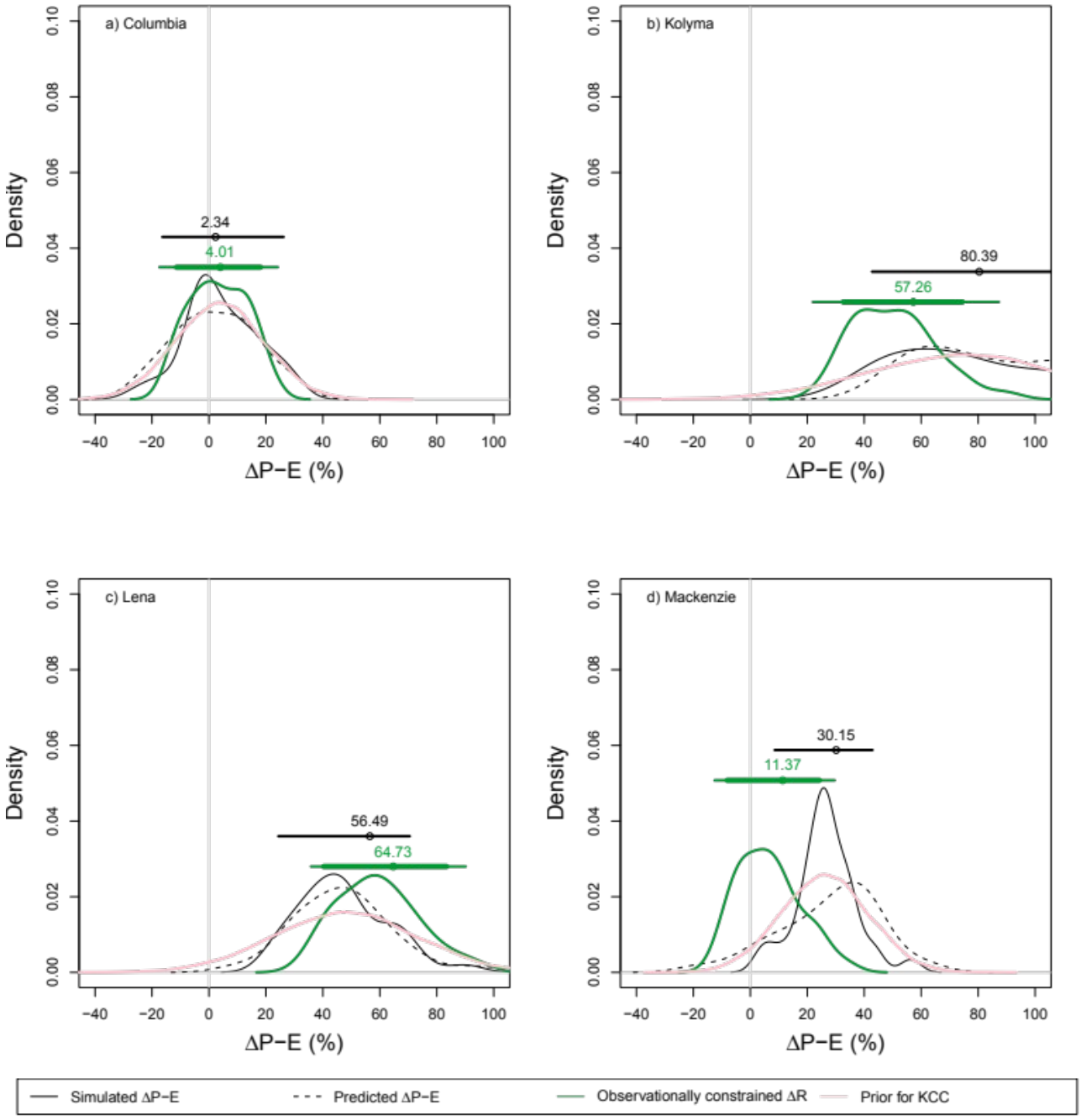

Figure S3: Constrained versus unconstrained distributions of water-year mean P-E relative anomalies (%) from the CMIP6 model ensemble under the SSP5-8.5 high-emission scenario: a) Columbia, b) Kolyma, c) Lena, and d) Mackenzie. Similar to Fig.1 but using P-E as a surrogate of R (although the regression coefficients are still constrained with GRUN runoff reconstructions).

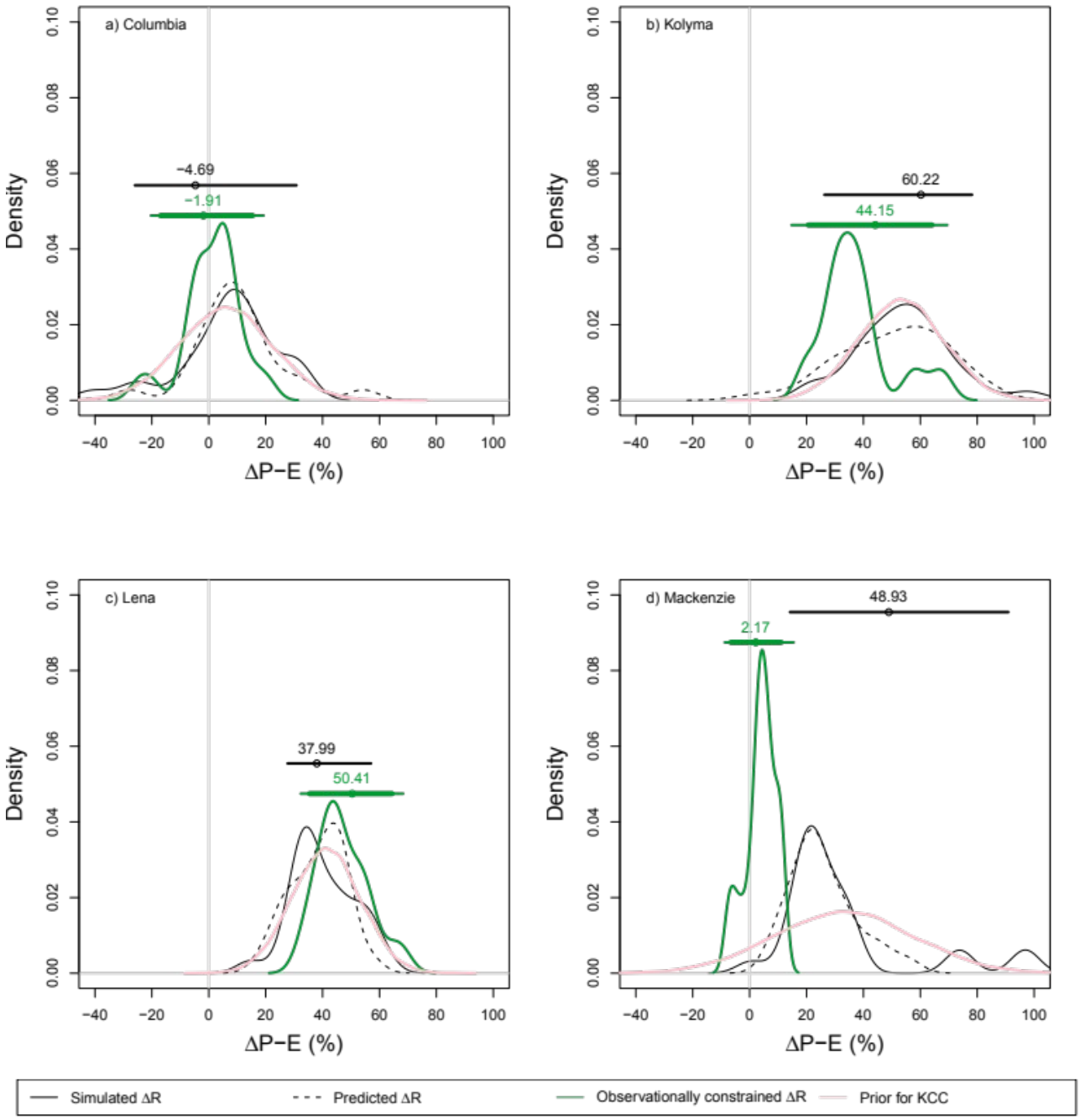

Figure S4: Constrained versus unconstrained P-E relative anomalies (%) from the CMIP5 model ensemble under the RCP8.5 high-emission scenario: a) Columbia, b) Kolyma, c) Lena, and d) Mackenzie. Similar to Fig.1 but using CMIP5 instead of CMIP6 models. The results are sensitive to the choice of the CMIP ensemble, although they are qualitatively consistent regarding the effect of the constraint on the ensemble mean.

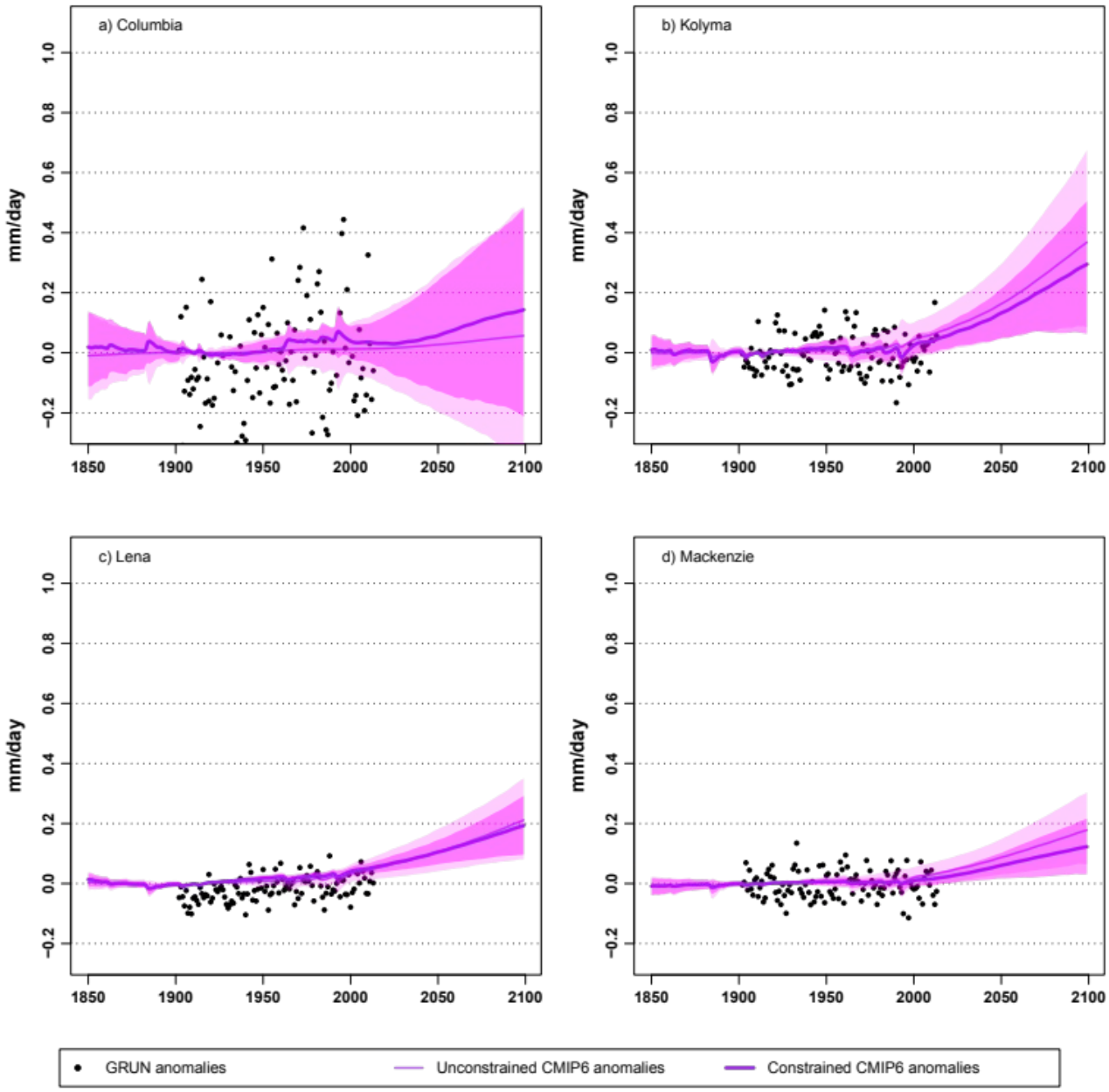

Figure S5: Constrained versus unconstrained water-year runoff anomalies (mm/day) using the rcp81 ensemble (a single realization of each CMIP5 model under the RCP8.5 scenario). Similar to Fig.2 but using the CMIP5 models. Depending on the CMIP model ensemble, the *prior* distribution is not the same but the effect of the KCC constraint on the *posterior* distribution is qualitatively consistent.

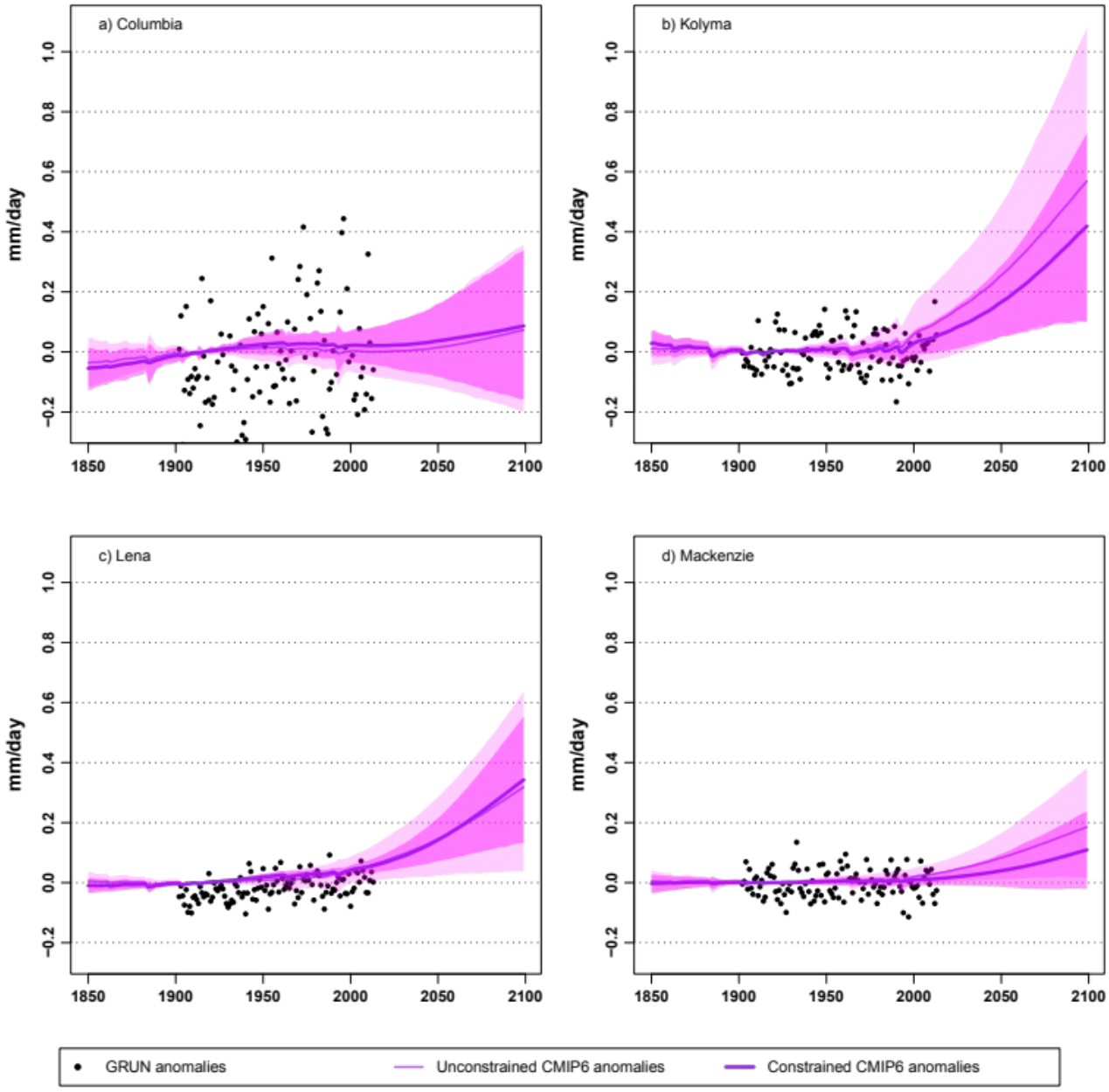

Figure S6: Constrained versus unconstrained water-year runoff anomalies (mm/day) using the ssp50 ensemble (multiple realizations of each CMIP6 model under the RCP8.5 scenario). Similar to Fig.2 but using multiple realizations for a lower number of CMIP6 models. Not surprisingly, differences between constraints using ssp50 and ssp51 scenarios are light, a smoothing of the distribution can be observed when more realizations are used (ssp50), less internal variability appears. There are no clear differences concerning the reduction of the spread.

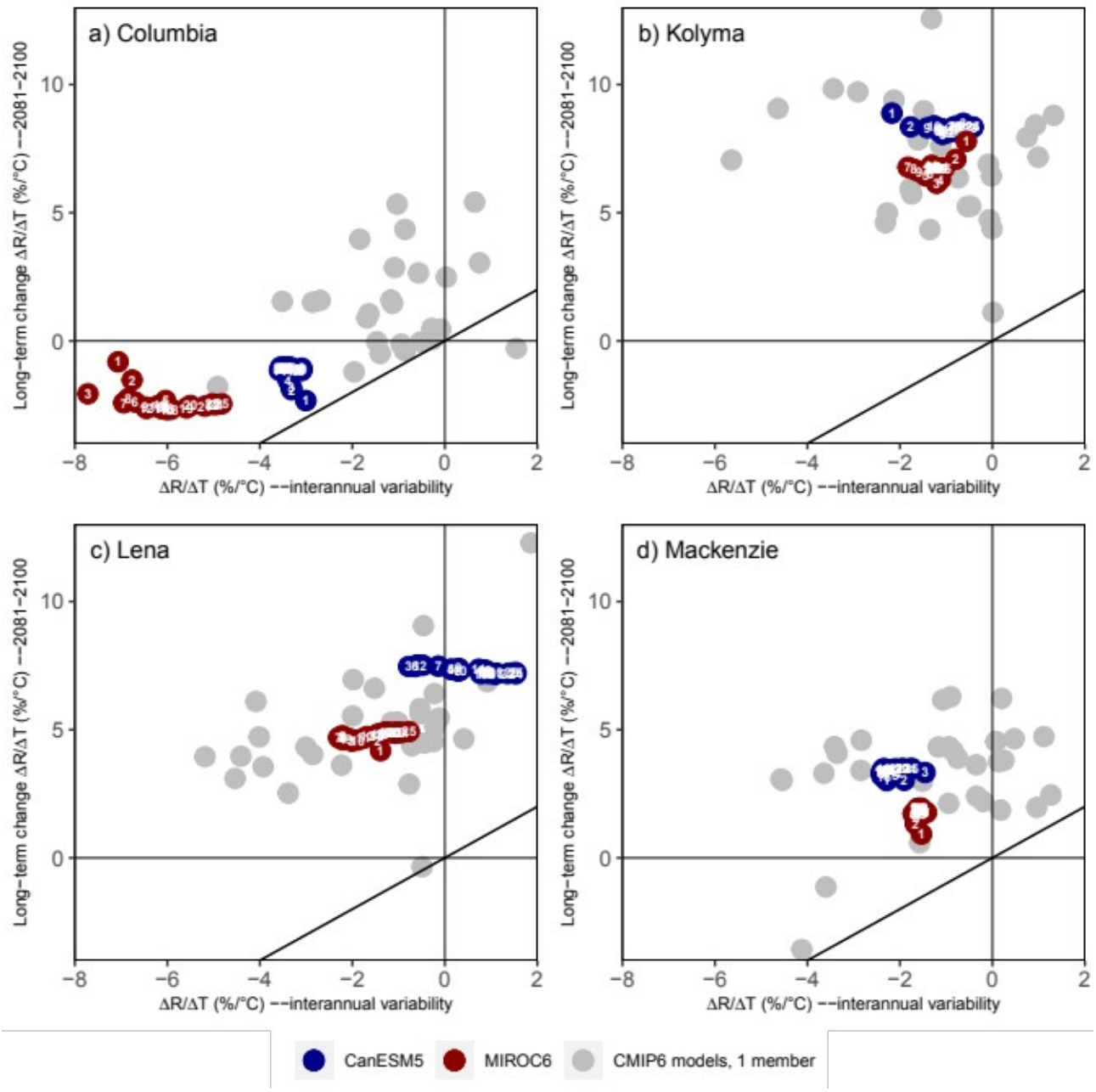

Figure S7: Runoff sensitivity to temperature at short (x-axis) versus long (y-axis) timescales. For each model (grey dots), the variation in runoff relative to the variation in temperature (%/°C) at the basin-scale is computed at two timescales. The long-term runoff sensitivity is estimated as the ratio of the averaged simulated anomalies over the 2081-2100 period compared to the 1902-1930 baseline. The short-term runoff sensitivity is estimated as the interannual variability over the 1902-2013 period (also used to estimate the regression coefficients in L19). The same computations are done for the CanESM5 and MIROC6 models (with 25 available realizations) after averaging an increasing number of realizations (for instance, dot 20 corresponds to average of the first 20 members). There is no obvious link between the runoff sensitivity to temperature at long versus short timescales. Therefore, the L19 hypothesis regarding the timescale independency of the runoff sensitivity to temperature does not seem valid.

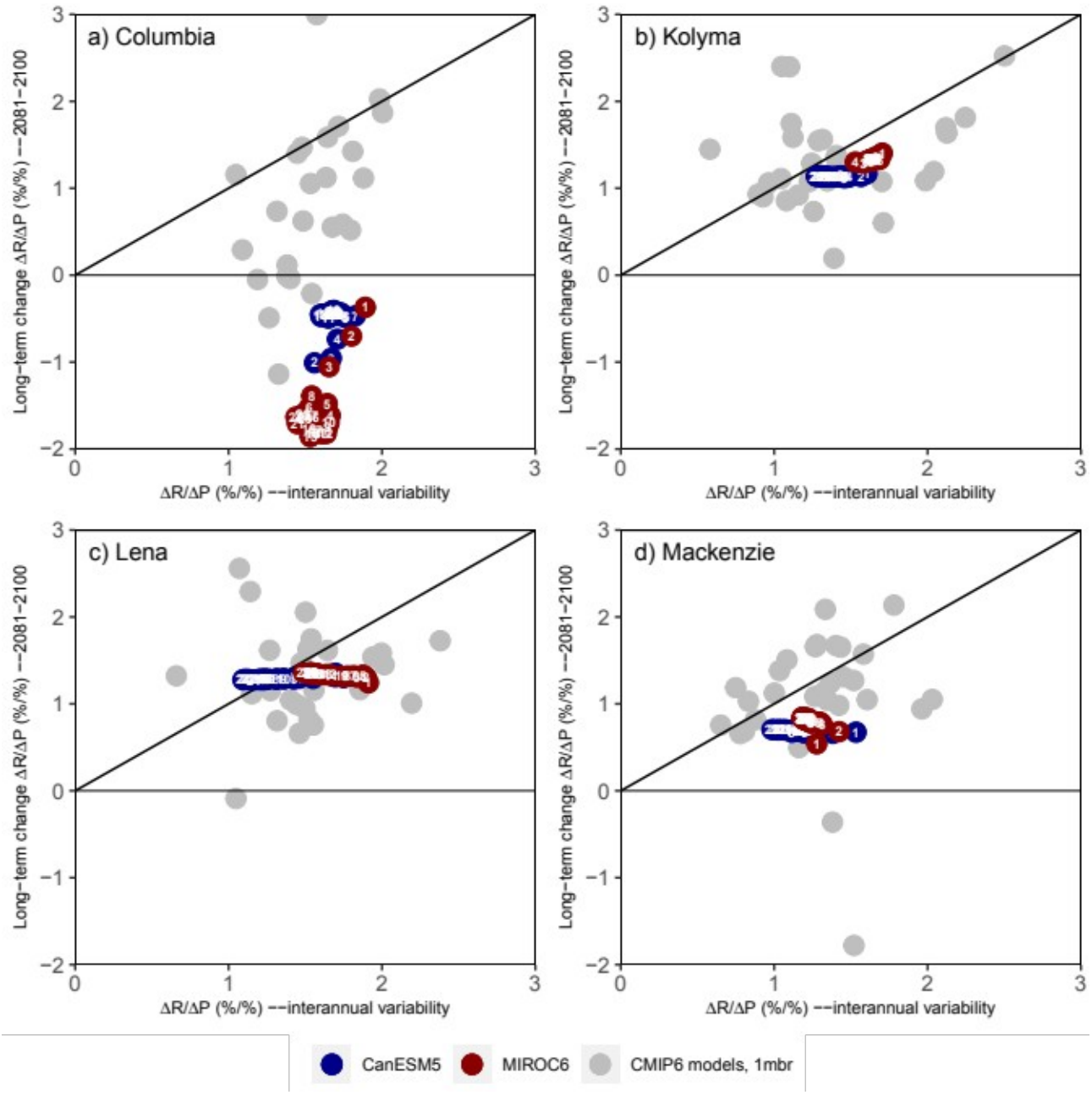

Figure S8: Runoff sensitivity to precipitation at short (x-axis) versus long (y-axis) timescales. Similar to Fig.S7 but comparing long and short-term runoff sensitivity to precipitation instead of temperature. Once again, the interannual runoff sensitivity is not a good surrogate for its long-term sensitivity to precipitation.