Uncorrected Proof

Journal of
Water & Climate Change

Journal of Water and Climate Change Vol 00 No 0, 1 doi: 10.2166/wcc.2021.345

Estimation of groundwater recharge using multiple climate models in Bayesian frameworks

Kevin O. Achieng[a,b,c,*] and Jianting Zhu[d]

[a] Department of Crop and Soil Science, University of Georgia, Athens, GA 30602, USA
[b] Department of Civil Engineering, Dedan Kimathi University of Technology, Private Bag 10143, Nyeri, Kenya
[c] Water Resource Management Center, Dedan Kimathi University of Technology, Nyeri, Kenya
[d] Department of Civil and Architectural Engineering, University of Wyoming, Laramie, WY 82071, USA
*Corresponding author. E-mail: kevin.achieng@uga.edu; kachiengz@yahoo.co.uk

KOA, 0000-0001-7815-9095

ABSTRACT

Groundwater recharge plays a vital role in replenishing aquifers, sustaining demand, and reducing adverse effects (e.g. land subsidence). In order to manage climate change-induced effects on groundwater dynamics, climate models are increasingly being used to predict current and future recharges. Even though there has been a number of hydrological studies that have averaged climate models' predictions in a Bayesian framework, few studies have been related to the groundwater recharge. In this study, groundwater recharge estimates from 10 regional climate models (RCMs) are averaged in 12 different Bayesian frameworks with variations of priors. A recession-curve-displacement method was used to compute recharge from measured streamflow data. Two basins of different sizes located in the same water resource region in the USA, the Cedar River Basin and the Rainy River Basin, are selected to illustrate the approach and conduct quantitative analysis. It has been shown that groundwater recharge prediction is affected by the Bayesian priors. The non-Empirical Bayes g-Local-based Bayesian priors result in posterior inclusion probability values that are consistent with the performance of the climate models outside the Bayesian framework. With the proper choice of priors, the Bayesian frameworks can produce good results of groundwater recharge with R^2, percent bias error, and Willmott's index of agreement of >0.97, $<2\%$, and >0.97, respectively, in the two basins. The Bayesian framework with an appropriate prior provides opportunity to estimate recharge from multiple climate models.

Key words: Bayesian framework, Bayesian priors (MPriors and g-priors), groundwater recharge, regional climate models (RCMs)

HIGHLIGHTS

- The choice of prior affects the suitability of Bayesian formulation for averaging recharge.
- All RCMs tend to underestimate groundwater recharge.
- Non-EBL-based priors result in posterior inclusion probabilities consistent with the RCM performance.
- Bayesian frameworks produce a better recharge estimate than individual RCMs.

1. INTRODUCTION

Globally, groundwater supplies 70% of water needs for irrigation (Rosegrant *et al.* 2009; Siebert *et al.* 2010; Long *et al.* 2020). About 25% of total freshwater consumption in the USA comes from groundwater (Meixner *et al.* 2016). Besides, groundwater is a source of safe drinking water to 90% of the US rural population (Niraula *et al.* 2017).

To assess the effect of potential climate change on groundwater resources, groundwater recharge is increasingly being modeled using climate models (Risser *et al.* 2005; Allen *et al.* 2010; Crosbie *et al.* 2010). The regional climate models (RCMs) use land surface models (LSMs) to partition precipitation from GCMs to surface runoff, evapotranspiration, and drainage (Mearns *et al.* 2012). These hydrological variables are often archived in databases such as North American Regional Climate Change Assessment Program (NARCCAP 2007; Mearns *et al.* 2013). Most LSMs are one-dimensional (1-D) vertical soil columns that partition precipitation to various hydrological outputs at relatively smaller spatial scales. Previous studies have successfully used RCMs and/or LSMs hydrological products to simulate the groundwater recharge. A Water Atmosphere Vegetation Energy and Solutes (WAVES) model (Zhang & Dawes 1998) is a 1-D soil water balance model and has been often and successfully used to simulate recharge by equating recharge to deep drainage below the rooting depth, assuming that soil evaporation and transpiration are negligible below the rooting depth (Crosbie *et al.* 2013; Xie *et al.* 2018). Hydrologic Evaluation of Landfill Performance (HELP) (Berger 2000, 2015) is a 1-D model, driven by daily precipitation, daily temperature, and daily solar radiation, which has been successfully used in numerous studies to simulate recharge based on deep drainage through and below the vertical soil column model (Woyshner & Yanful 1995; Berger 2000, 2015; Risser *et al.* 2005; Scibek & Allen 2006; Jyrkama & Sykes 2007; Allen *et al.* 2010). Soil Vegetation Atmosphere Transfer (SVAT) is a class of 1-D LSM that takes rainfall as input and partitions it into ET, runoff, and drainage below the soil column (recharge). SVAT has been used in multiple recharge studies such as in the Okavango River Basin and the Chobe-Zambezi River system (Brunner *et al.* 2004).

Combining multiple climate models' hydrological predictions in a Bayesian framework (Sloughter *et al.* 2007; Duan & Phillips 2010) has been found to provide better predictions than the individual climate models (Ma *et al.* 2016). Some of the hydrological studies based on the Bayesian framework include: the streamflow study on Mississippi's Leaf River Basin (Ajami *et al.* 2007) and French Broad watersheds in the USA (Vrugt *et al.* 2008); Yellow River Basin, China (Zhang *et al.* 2018); Irish river catchments (Bastola *et al.* 2011); and Bayesian-based precipitation and sea surface temperature studies (Luo *et al.* 2007). However, there are limited recharge studies, if any, that have averaged recharge prediction from multiple climate models in a Bayesian framework. In addition, most previous hydrological studies were based on an assumed single Bayesian prior. However, due to subjectivity, an assumed single Bayesian prior may have a potential of an ill-posed posterior model probability (PMP; Kavetski *et al.* 2006) potentially due to drastic changes in posterior probability because of small errors in the observation. Bayesian model averaging (BMA) has been used to quantify groundwater model uncertainty, with model parameters being assumed to follow a uniform prior distribution (Mustafa *et al.* 2020). In a study conducted in the San Joaquin River Basin, BMA was successfully used to estimate the uncertainty of aquifer storage from the machine leaning-based groundwater model (Yin *et al.* 2021). Plausible recharge estimates were obtained in Central America using a combination of stable isotopes and the BMA framework with an assumed uniform prior distribution (Arellano *et al.* 2020). Groundwater storage across the contiguous USA was reliably estimated using a combination of Gravity Recovery and Climate Experiment (GRACE) satellite data and the water balance model within the Bayesian framework (Mehrnegar *et al.* 2021). The Bayesian framework, with an assumed uniform distribution, was used to establish interactions of climate with groundwater and food within Louisiana, USA (Singh *et al.* 2020). Groundwater age-dating within the Bayesian framework (with an assumed log-normal prior distribution) was used to better understand the irregular mixing at the seawater intrusion zones of groundwater aquifer that exist along the shorelines of Yellow Sea, South Korea (Ju *et al.* 2021). In a study conducted in the Yixunhe watershed, China, non-point source pollution was reliably modeled in the BMA framework (Wang *et al.* 2020). Other studies have also investigated Bayesian-based integrated management of stressed hydrological systems (Molina *et al.* 2010) and the impact of climate change on stressed groundwater (Molina *et al.* 2013). Therefore, there is no study that has investigated the influence of Bayesian priors on groundwater recharge estimation to the best of our knowledge.

Based on the above discussion, the objectives of this study are to (1) average recharge predictions from 10 RCMs in 12 different Bayesian frameworks; (2) find a suitable Bayesian framework for averaging multiple climate-based recharge prediction; (3) evaluate the relative importance of climate models in predicting recharge inside and outside the Bayesian framework. To calculate posterior model probabilities, which are used to average recharge prediction from the climate models, a known actual/observed recharge is required. In this study, a recession-curve-displacement method, which is also called RORA method (Rorabaugh 1964; Rorabaugh & Simons 1966), is used to estimate groundwater recharge from historical daily streamflow measurement. The RORA method has been proven to provide reliable recharge estimates from historical streamflow measurements (Daniel 1976; Delin *et al.* 2007; Lorenz & Delin 2007; Rutledge 2007).

2. METHODS

2.1. Study area and data sources

In this study, two case-study basins located in Minnesota, USA are selected. The Cedar River Basin (near Austin, Minnesota) and the Rainy River Basin at Manitou Rapids have drainage area of 1,500 and 30,000 km^2, respectively. The gauging station for the Cedar River Basin is located at [−92.97°, 43.64°], whereas for the Rainy River Basin it is located at [−95.54°, 44.72°]. The Cedar Basin receives relatively higher rainfall (of 834 mm/year) than the Rainy Basin which receives 671 mm/year on average (Achieng & Zhu 2019).

The estimation of actual groundwater recharge using the RORA method requires continuous data of historical streamflow measurement. Both Cedar and Rainy River Basins' streamflow dataset was obtained from the United States Geological Survey (USGS). The climate model's water balance dataset is freely available from the North American Regional Climate Change Assessment Program (NARCCAP) (Mearns *et al.* 2012). The NARCCAP project has downscaled the general circulation models (GCMs) to the RCMs' relatively smaller spatial and temporal resolutions of 50 km and 3 h, respectively. The RCMs have LSMs that partition the precipitation input into various hydrological outputs at RCMs' spatial and temporal scales. Comparison of the climate models requires use of a common time-frame; therefore, the RCM datasets of 1968–1998 are used in this study. The same 10 NARCCAP RCM–GCM combinations as the study of Achieng & Zhu (2019) are used in this study. It should be noted, however, that any number of RCMs can be included in the proposed framework. Table 1 summarizes the RCMs and their corresponding LSMs.

For the RCMs, CRCM is the Canadian RCM (Caya & Laprise 1999; Music & Caya 2007), ECP2 is the Scripps Experimental Climate Prediction Center/Regional Spectral Model (Juang *et al.* 1997), MM5I is the Fifth-generation Pennsylvania State University /NCAR Mesoscale Model (Grell *et al.* 1994), RCM3 is the RCM, version 3 (Pal *et al.* 2007), and WRFG is the Weather Research and Forecasting Grell (Skamarock *et al.* 2005) model. For the GCMs, ccsm is the Community Climate System Model (Collins *et al.* 2006), cgcm3 is the Third Generation Coupled Global Climate Model (Flato *et al.* 2000), gfdl is the Geophysical Fluid Dynamics Laboratory (GFDL) model (Delworth *et al.* 2006), and hadcm3 is the Hadley Climate Model (Pope *et al.* 2000).

2.2. RCM-based recharge estimate

Since the RCMs use the 1-D vertical column LSMs to partition precipitation from GCMs, the drainage below the soil column is assumed to be recharge based on the following three assumptions: (1) the downward flux is not blocked by impermeable layer, and thus the lateral flow is assumed to be negligible, and (2) the lag-time between drainage and recharge is relatively short (i.e. hours to a few days to reduce losses such as evapotranspiration – which would reduce the amount of recharge), and (3) climate and land use/land cover do not change during the period when drainage leaves the bottom of the soil column and when drainage reaches the water table. RCMs' LSMs have been used to estimate recharge from point scales (Zhang & Dawes 1998; Crosbie *et al.* 2013) to regional scales (Risser *et al.* 2005; Scibek & Allen 2006; Allen *et al.* 2010). Based on these assumptions, the water balance for the 10 NRCCAP RCMs (i.e. CRCM_ccsm, CRCM_cgcm3, ECP2_gfdl, ECP2_hadcm3, MM5I_ccsm, MM5I_hadcm3, WRFG_ccsm, and WRFG_cgcm3) is expressed as:

$$R = [pr + snm] - evps - mrros \pm \Delta S \tag{1}$$

where R is the recharge (mm), pr is the rainfall (mm), snm is the snowmelt (mm), $evps$ is the evapotranspiration (mm); $mrros$ is the surface runoff (mm), and ΔS is the change in water storage (mm). The NARCCAP models provide the water balance

Table 1 | NARCCAP RCMs and corresponding LSMs

NARCCAP RCMs	LSMs
1. CRCM_ccsm	Canadian Land Surface Scheme (CLASS) (Verseghy; Verseghy *et al.* 1993).
2. CRCM_cgcm3	
3. ECP2_ gfdl	NOAH (N: National Centers for Environmental Prediction (NCEP); O: Oregon State University (Department of Atmospheric Sciences); A: Air Force (both Air Force Weather Agency and Air Force Research Laboratory); H: Hydrology Lab – National Weather Service, NWS) (Ek *et al.* 2003).
4. ECP2_hadcm3	
5. MM5I_ccsm	
6. MM5I_hadcm3	
7. WRFG_ccsm	
8. WRFG_cgcm3	
9. RCM3_cgcm3	Biosphere-Atmosphere Transfer Scheme (BATS) (Dickinson *et al.* 1986; Yang & Dickinson 1996).
10. RCM3_gfdl	

variables on 3-hourly intervals. The 3-hourly values are summed up into monthly values because recharge values are at monthly time-steps. Therefore, the terms in Equation (1) are on a monthly time-step. Note that in 2 of the 10 RCMs, RCM3_ cgcm3 and RCM3_ gfdl, the rainfall and snowmelt were combined as one variable.

To calculate the average recharge for the basin, recharge estimates from the RCM grid cells that cover the basin are averaged using the grid cells areas as the weights. Since the studied basins have irregular shapes and the RCM grids are regular in shape, only the portions of the RCM grids that cover the basins are considered.

2.3. Observed groundwater recharge estimation: recession-curve-displacement method

The climate models' recharges are averaged in a Bayesian framework based on a known actual recharge. The observed recharge is estimated from the USGS continuous historical daily streamflow measurements using the RORA method (Rorabaugh *et al.* 1966; Rorabaugh 1964). The RORA method assumes that recharge events only occur in the aquifer and instantaneously cause the groundwater level to rise, while the rest of the basin does not experience any net gain or loss. Therefore, the rise in the groundwater level causes groundwater discharge into the stream. The Cedar Basin has sand/gravel aquifer that is about 30 m thick. The Rainy Basin has igneous and metamorphic rock-based aquifer that stores water in fractures and faults. The Cedar Basin is most likely to uphold the homogeneity/isotropy assumption because its aquifer is made of coarse sediment. Since the Rainy Basin's aquifer is composed of rock fractures and faults, it may violate the homogeneity/isotropy assumption if these cracks are not uniformly distributed throughout the aquifer. This assumption is appropriate for the two basins because both basins experience humid climate. Therefore, it is possible for the basins to have zero net loss (e.g. due to evapotranspiration) or net gain (e.g. from irrigation). The time period between the peak streamflow discharge and the stabilized streamflow is called the critical time. It is used in this study because it has been found to provide groundwater recharge that is representative of the entire river basin (Delin *et al.* 2007; Lorenz & Delin 2007) unlike the point-scale methods. Besides, the RORA method only needs streamflow dataset in order to determine the recharge (Busenberg & Plummer 1992).

A recession index that is related to the aquifer hydrogeological properties is determined as the time interval after the critical time, which is required for the groundwater discharge to recede by 1 log-cycle. The recession index is computed from the recession curve, which is a semi-log plot of groundwater discharge after the critical time versus time. After identifying the peak discharge and the critical time, the discharge at the start of the critical time is extrapolated to give the pre-event and that at the end of the critical time is extrapolated to give the post-event discharge. Both the recession index and the critical time can be computed as a function of distance between the stream and the groundwater divide, aquifer transmissivity, and aquifer storativity (Rorabaugh 1964; Delin *et al.* 2007). Groundwater recharge associated with each peak event is analytically calculated from the recession-curve-displacement as function of the difference between the pre-event and post-event discharges as follows (Rorabaugh 1964; Rorabaugh & Simons 1966; Daniel 1976; Arnold & Allen 1999; Rutledge 2000, 2004; Delin *et al.* 2007; Lorenz & Delin 2007):

$$R = \frac{2(Q_2 - Q_1)}{2.3026A}\,K = \frac{(Q_2 - Q_1)}{1.1513A}K \tag{2}$$

where R is the daily groundwater recharge (mm); A is the area of the basin (km^2); Q_1 is the groundwater discharge at the critical time that is extrapolated from the pre-event streamflow recession (m^3/s); Q_2 is the groundwater discharge at the critical time that is extrapolated from the post-event streamflow recession (m^3/s); K is the recession index (days/log-cycle) which is either obtained as the slope of the recession curve after critical time or computed as a function of aquifer properties as shown in the Appendix.

2.4. Bayesian frameworks for averaging RCMs' recharge

Regression models are developed between the actual recharge and the RCMs' recharge in the 2^k model space, where k is the total number of regressors (RCMs), i.e. 10. The actual recharge Y_O is regressed onto each of the possible combinations of the 10 regressors to produce a total of 2^{10} (i.e. 1,024) regression models. A normal regression model M_j of actual recharge onto k_j regressors (where $0 \le k_j \le k$) that contains n observations is expressed as:

$$Y_O = I_n\mu + Z_j\beta_j + \sigma\varepsilon \tag{3}$$

where μ is the intercept, I_n is the n-dimensional vector of 1's, Z_j is an $n \times k_j$ matrix that contains k_j columns of regressors and n rows of observations of the respective regressors, β_j is the k_j-dimensional vector of relevant regression coefficients, σ is the residual standard deviation, ε is the identically and normally distributed ratio of residual to error that results from fitting all the k regressors to Y_O.

Twelve Bayesian frameworks, which are used in the study by Achieng & Zhu (2019), were implemented as the Bayesian recharge models. These Bayesian frameworks are developed by combining prior distribution of the Bayesian regression models (MPrior) and the prior distribution of the regression coefficients (g-prior). Both prior distributions are vital in computing PMP, which is then used to compute both the Bayesian regression coefficients and the posterior inclusion probability (PIP) of the individual covariates (i.e. the RCMs).

The PMP of the regression model j is computed, based on Bayes theorem, as follows:

$$P(M_j|Y_O) = \frac{P(Y_O|M_j) \times P(M_j)}{\sum_{h=1}^{2^k} P(Y_O|M_h) \times P(M_h)} \tag{4}$$

where $P(Y_O|M_j)$ is the likelihood of regression model M_j that is computed based on the g-prior, $\sum_{h=1}^{2^k} P(Y_O|M_h) \times P(M_h)$ is the normalizing constant of the PMP, $P(M_j)$ is the prior probability distribution of the regression model M_j, which is the MPrior, Y_O is the actual recharge, and k is the number of RCMs, i.e. 10.

The PIP of regressor (i.e. RCM) j is computed as the sum of PMPs of regression models that contain the regressor j:

$$\mathrm{PIP} = P(\beta_j \neq 0|Y_O) = \sum_{\beta_j \neq 0} P(M_j|Y_O) \tag{5}$$

where β_j is the regression coefficient. PIP is used because it has been found to be useful performance indicator in evaluating the relative importance of the explanatory variables within the Bayesian framework (Fernández *et al.* 2002; Zeugner 2011; Ley *et al.* 2012; Steel 2013; Zeugner & Feldkircher 2015).

The likelihood of regression model is computed by integrating, with respect to the regression coefficients, the joint probability density functions of the regression models, and regression coefficients (μ, σ^2, β_j) as (Fernández *et al.* 2001a, 2001b):

$$P(Y_O|M_j) = \int\int\int P(Y_O|\mu, \beta_j, \sigma^2, M_j)P(\mu, \sigma^2)P(\beta_j|\mu, \sigma^2, M_j)d\mu d\sigma^2 d\beta_j \tag{6}$$

Three MPriors were used in this study including (1) the discrete uniform distribution (Liang *et al.* 2008; Ley & Steel 2012), (2) the fixed model prior (Ley & Steel 2008), and (3) the random model prior (Zeugner 2011; Zeugner & Feldkircher 2015). Along with the three MPriors, four g-priors were used in this study which includes: (1) the uniform information prior (UIP) (Kass & Wasserman 1995), (2) the Bayesian Risk Information Criterion (BRIC) (Fernández *et al.* 2001a), (3) hyper-UIP (Liang *et al.* 2008), and (4) the Empirical Bayes g-Local (EBL) (George & Foster 2000; Cui & George 2004) g-priors. A total of 12 Bayesian frameworks were formulated based on these MPriors and g-priors. Further details about these Bayesian frameworks can be found in Achieng & Zhu (2019).

PMP is used to compute the Bayesian regression coefficient as follows:

$$E(\beta_j|M_j,\ Z) = \sum_{j=1}^{2^k} E(\beta_j|Y_O, M_j,\ Z) \times P(M_j|Y_O) \tag{7}$$

where $E(\beta_j|M_j,\ Z)$ is the posterior mean expected value of the Bayesian regression coefficient, $E(\beta_j|Y_O, M_j,\ Z)$ is the estimated coefficient β_j from the model M_j, $P(M_j|Y_O)$ is the PMP of the model j, and $Z \in \{x_1,\ x_2,\ \ldots,\ x_k\}$ is the covariates matrix. Using Equation (7), the Bayesian averaged coefficients are computed for each of the k regressors. The Bayesian recharge model is comprised of these k regressors with their corresponding computed regression coefficients.

2.5. Evaluation of the performance of the climate models and Bayesian recharge models

The evaluation of the performance of the climate models and the Bayesian frameworks is performed to understand the strength of the respective models in simulating recharge, inform parameter tuning of the climate models, and evaluate the advancement of the current versions of the climate models compared with the previous versions. In this study, four widely used classical performance indices are used to assess the performance of the RCMs and the Bayesian recharge models: the coefficient of determination (R^2) (Gupta *et al.* 1998; Dawson *et al.* 2002), the root mean square error (RMSE) (Legates & McCabe 1999), the Willmott's index of agreement (d1) (Willmott 1982; Willmott *et al.* 1985), and the percent bias error (PBIAS) (Gupta *et al.* 1998). The RMSE is used because it is a measure of goodness of fit between observation and modeled recharge values. The PBIAS gives us the deviation of the modeled recharge values from the observation. The index of agreement d1 characterizes how well the modeled recharge is in agreement with the observation.

The relative importance of the 10 climate models is evaluated both before (outside the Bayesian framework) and after (inside the Bayesian framework) the climate models are averaged in a Bayesian framework. Evaluating the relative importance outside the Bayesian framework is done to ensure that Bayesian prior probability results into PMP values that are consistent with the performance indices of the climate models that are obtained outside the Bayesian framework (Achieng *et al.* 2019).

2.6. Sensitivity analysis of Bayesian recharge models

Bayesian recharge models are formulated by regressing the actual recharge on to the 10 RCMs' predicted recharge. Sensitivity of these recharge models, to the covariates, is assessed based on a leave-one-out approach. A climate model is left out of the trained Bayesian recharge regression equation, and the recharge prediction of the resulting model is compared with the prediction of the full model. The magnitude and direction of the change in recharge prediction, when one of the climate models is left out, give an indication of the sensitivity of the trained Bayesian recharge model to the climate model in question. The climate model which leads to the greatest change in recharge is the most important model in the Bayesian regression model. Since averaging of the climate models within the Bayesian framework aims to use the relative importance of the climate models in predicting the Bayesian-based recharge, the climate model that results in a relatively big change in predicted recharge when left out of the Bayesian recharge model is the climate model that the recharge model is most sensitive to.

3. RESULTS AND DISCUSSION

3.1. Historical recharge

The results of historical annual recharge for both basins are shown in Figure 1. Recharge estimation with this method involves the determination of both the recession index and the critical time for each basin. The recession index values for Cedar and Rainy Basins were 19.78 and 73.94 days/log-cycle, respectively. Whereas the critical time values for Cedar and Rainy Basins were 4.24 and 15.9 days, respectively. The Cedar River Basin seems to have a positive trend in annual recharge (of 3.4 mm/year) over the 30 years, whereas the Rainy Basin has a negative trend in annual recharge (of −3.4 mm/year) over the same period. This signifies an increase and a decline in annual recharge in Cedar and Rainy River Basins, respectively, during the 1968–1998 time period. Note that the long trend in an annual recharge observed in these basins is fairly weak since both basins have a low R^2 term value of 0.15. The Rainy River Basin has a long-term mean annual recharge of 225 mm, which is larger than that of the Cedar River Basin of 191 mm. The relatively larger mean annual recharge in the Rainy River Basin can be explained by the fact that this basin lies in the northern Minnesota – a region that receives relatively higher precipitation than the southern region where the Cedar River Basin is located. These recharge values are similar those obtained in previous studies (Delin *et al.* 2007; Lorenz *et al.* 2007).

3.2. Recharge estimation from climate models

The scatterplots of monthly recharge for each RCM and observed data are plotted as shown in Figure 2 for both Cedar and Rainy River Basins. Most RCMs (e.g. WRFG_ccsm, WRFG_cgcm3, and MM5I_ccsm) underestimated recharge, whereas a few RCMs (e.g. ECP2_hadcm3and RCM3_cgcm3) overestimated recharge in both basins. Due to variability in partitioning of the precipitation, the RCMs' LSMs provide different values of the water balance components. As a result, there exists variability in recharge prediction among the RCMs as shown in Figure 2. For example, studies have found that NOAH LSM that is used by most of the NARCCAP RCMs (such as ECP2_ gfdl, ECP2_hadcm3, MM5I_ccsm, MM5I_hadcm3, WRFG_ccsm, and WRFG_cgcm3) underestimates evapotranspiration (Yang *et al.* 2011) and overestimates surface runoff (Niu *et al.* 2011) because the frozen top soil layer in the NOAH model is relatively more impermeable than its counterpart LSMs, thus blocking infiltration and deep drainage while promoting soil evaporation. Other studies have found out that the Canadian Land Surface Scheme (CLASS) (which is used by CRCM_ccsm and CRCM_cgcm3 climate models) has been found to

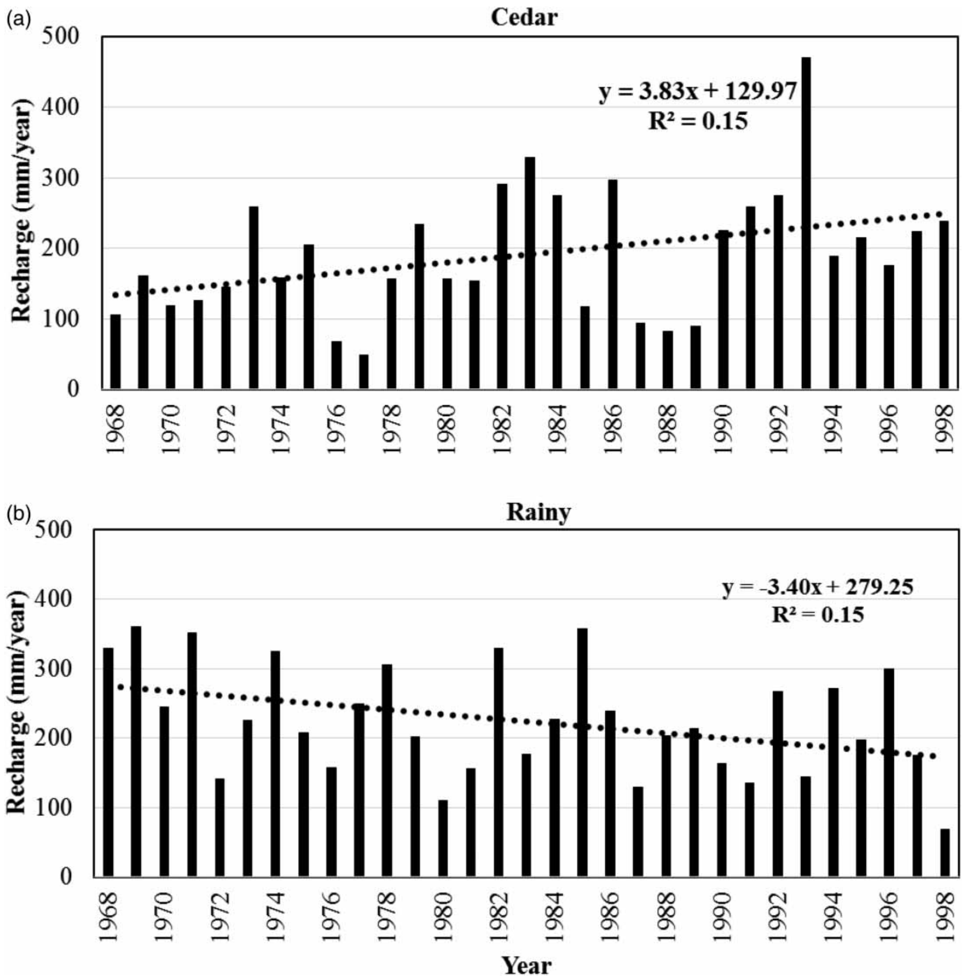

Figure 1 | Historical annual groundwater recharge in (a) Cedar and (b) Rainy Basins.

underestimate surface runoff and drainage (Lohmann *et al.* 1998) and overestimates evapotranspiration (Liang *et al.* 1998). On the other hand, the Biosphere-Atmosphere Transfer Scheme (BATS) is prone to overestimating surface runoff (Yang *et al.* 2011) and underestimating evapotranspiration (Dickinson & Henderson-Sellers 1988). The BATS is used by RCM3_cgcm3 and RCM3_gfdl RCMs.

3.3. Performance of the individual RCMs

Table 2 summarizes the performance indices of climate models in simulating the groundwater recharge. In the Cedar River Basin, RCM3_gfdl is the best performing RCM with R^2, RMSE, d1, and PBIAS of 0.98, 3.12 mm, 0.92, and 0.29%, respectively. In the Rainy River Basin, RCM3_cgcm3 is the best performing RCM with R^2, RMSE, d1, and PBIAS of 0.98, 4.04 mm, 0.93, and −8.79%, respectively. In both basins, CRCM_ccsm is the worst performing model. The poor performance of the CRCM_ccsm model can be attributed to the fact that it uses the CLASS LSM which has been found to overestimate both evapotranspiration (Yang *et al.* 2011) and surface runoff (Niu *et al.* 2011) and therefore to underestimate aquifer recharge. Both evapotranspiration and surface runoff are key components of the water balance-based recharge estimation. The value of R^2 of the remaining RCMs ranges from 0.55 to 0.98, for the Cedar Basin, and 0.22 to 0.99, for the Rainy Basin. The bias of the remaining climate models, which is represented with RMSE, is in the range of 4.5–19.1 mm in the Cedar River Basin and 4.0–35.6 mm in the Rainy River Basin. The remaining RCMs predicted recharge with d1 values of 0.50–0.92 and 0.26–0.90 in Cedar and Rainy River Basins, respectively. Whereas the PBIAS values with which the remaining RCMs predicted recharge range from −7.41 to +1.09% in the Cedar Basin and −14.77 to −1.13% in the Rainy Basin. The RCMs driven by 'gfdl' and 'cgcm3' GCMs (e.g. RCM3_cgcm3, RCM3_gfdl, and ECP2_gfdl) seem to perform better in

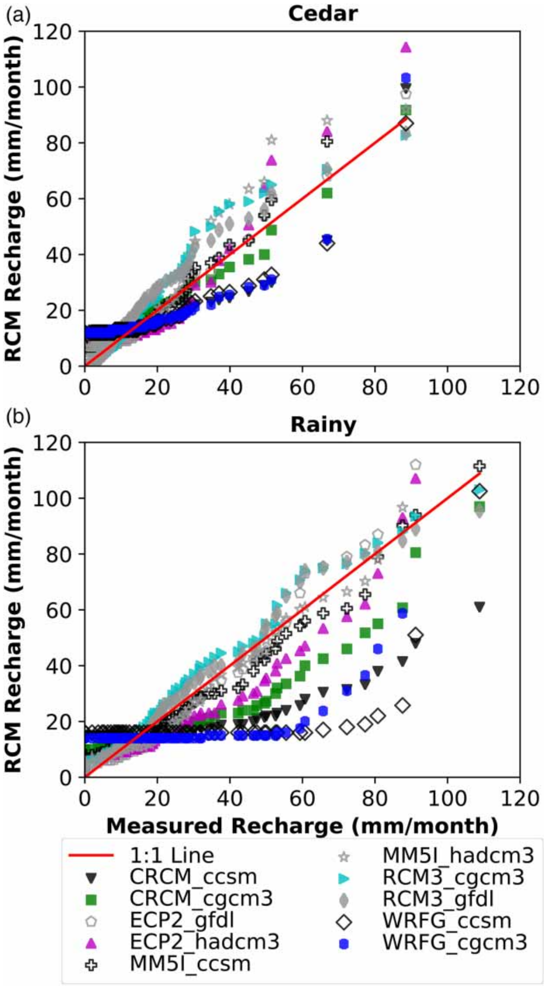

Figure 2 | RCM predicted recharge for (a) Cedar and (b) Rainy Basins for 1968–1998.

Table 2 | Performance of the climate models at simulating groundwater recharge

	Cedar				**Rainy**			
Climate models	R^2 **(–)**	**PBIAS (%)**	**d1 (–)**	**RMSE (mm)**	R^2 **(–)**	**PBIAS (%)**	**d1 (–)**	**RMSE (mm)**
CRCM_ccsm	0.552	−7.41	0.539	19.1	0.27	−14.8	0.353	43.2
CRCM_cgcm3	0.849	−3.48	0.841	9.64	0.46	−9.99	0.615	32.1
ECP2_gfdl	0.984	−0.82	0.921	2.61	0.91	−2.23	0.902	9.42
ECP2_hadcm3	0.879	−1.10	0.743	7.29	0.82	−1.13	0.814	12.9
MM5I_ccsm	0.888	−0.73	0.823	7.35	0.67	−6.50	0.776	22.2
MM5I_hadcm3	0.949	−0.14	0.848	4.52	0.86	−2.77	0.826	11.5
RCM3_cgcm3	0.924	1.09	0.887	5.52	0.99	−8.79	0.928	4.05
RCM3_gfdl	0.976	0.29	0.921	3.13	0.99	−7.82	0.939	3.58
WRFG_ccsm	0.569	−4.44	0.536	16.7	0.22	−6.79	0.266	35.6
WRFG_cgcm3	0.524	−1.66	0.504	16.6	0.37	−4.96	0.415	29.5

estimating recharge in both basins. On the other hand, the RCMs driven by 'ccsm' GCMs (such as CRCM_ccsm and WRFG_ccsm) have poor performance in recharge estimation.

3.4. Calibration/training of the Bayesian recharge models

The calibration of the Bayesian regression model is done by determining the optimal values of regression coefficients of each of the 10 regressors. Each of the regression coefficient (of the calibrated Bayesian model) is obtained by averaging the values of the regression coefficient of all the regression models that contain the regressor in question (within the 1,024 regression model space) using the corresponding PMP values of the regression models. A total of 12 Bayesian recharge models are calibrated using a subset of 23-year data from the 31-year dataset from 1968 to 1990. Calibration is undertaken using monthly data for the 1968–1990 period. Using the calibration dataset, the recharge predictions from the 10 climate models are averaged in the Bayesian framework with the RORA-based recharge being used as the observed recharge. Once the optimal Bayesian regression coefficients for the 10 RCMs are obtained, the calibrated Bayesian recharge model predictions are plotted versus the measured recharge as shown in Figure 3. Results suggest a strong correlation between the calibrated recharge models and the measured recharge. Nearly all the Bayesian recharge models have good performance with an R^2 value of 0.998, an RMSE value of <2 mm, d1 of >0.97, and PBIAS of <|1%| in both Cedar and Rainy River Basins, as shown in

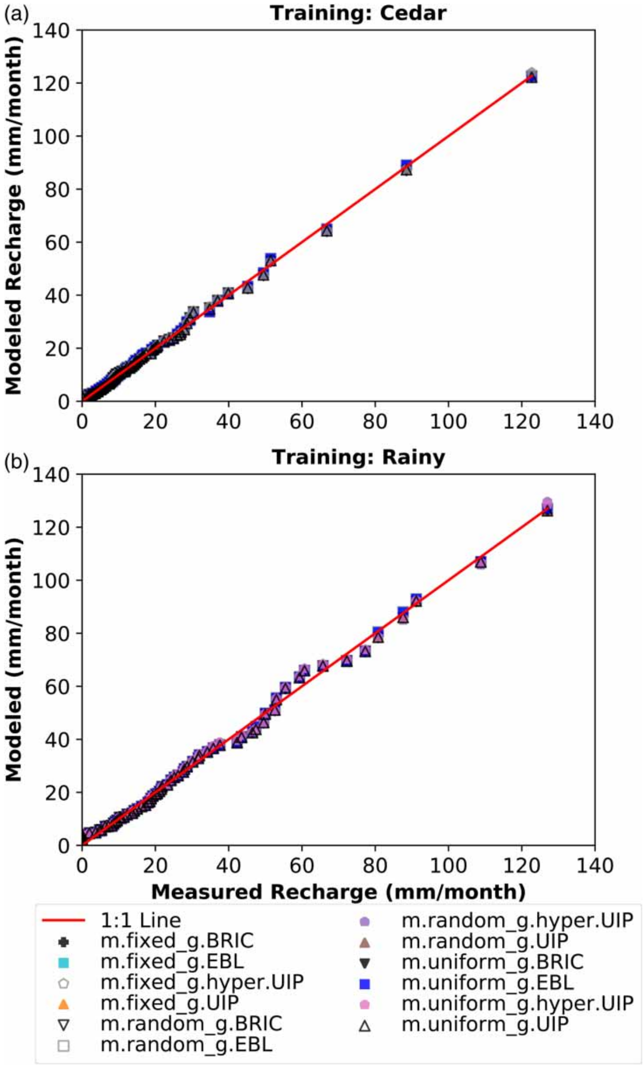

Figure 3 | Trained Bayesian models for (a) Cedar and (b) Rainy Basins using (**1968–1990**) data.

Table 3 | Performance of the trained Bayesian recharge models (using **1968–1990** calibration data)

	Cedar				**Rainy**			
Bayesian models	***R*² (–)**	**RMSE (mm)**	**d1 (–)**	**PBIAS (%)**	***R*² (–)**	**RMSE (mm)**	**d1 (–)**	**PBIAS (%)**
m.fixed_g.BRIC	0.998	0.858	0.974	-2.06×10^{-10}	0.996	1.64	0.974	−0.169
m.fixed_g.EBL	0.998	0.749	0.977	-2.75×10^{-10}	0.996	1.60	0.974	−0.363
m.fixed_g.hyper.UIP	0.997	0.926	0.970	-1.03×10^{-9}	0.996	1.66	0.973	−0.595
m.fixed_g.UIP	0.998	0.858	0.974	3.58×10^{-9}	0.996	1.64	0.974	−0.169
m.random_g.BRIC	0.998	0.898	0.972	2.06×10^{-9}	0.996	1.65	0.973	−0.195
m.random_g.EBL	0.998	0.749	0.977	1.10×10^{-9}	0.996	1.61	0.974	−0.364
m.random_g.hyper.UIP	0.998	0.898	0.972	-1.10×10^{-9}	0.996	1.71	0.972	−0.703
m.random_g.UIP	0.998	0.898	0.972	-1.10×10^{-9}	0.996	1.65	0.973	−0.195
m.uniform_g.BRIC	0.998	0.858	0.974	1.45×10^{-9}	0.996	1.64	0.974	−0.169
m.uniform_g.EBL	0.998	0.749	0.977	-1.45×10^{-9}	0.996	1.60	0.974	−0.363
m.uniform_g.hyper.UIP	0.997	0.926	0.970	3.92×10^{-9}	0.996	1.66	0.973	−0.595
m.uniform_g.UIP	0.998	0.858	0.974	2.68×10^{-9}	0.996	1.64	0.974	−0.169

Table 3. The monthly recharge of the remaining 2-year dataset for 1991–1998 is used to validate the 12 Bayesian recharge models.

3.5. Validation of the Bayesian recharge models

To ensure that the calibration of the Bayesian recharge models works effectively, the calibrated Bayesian recharge models were validated. The monthly recharge of the 1991–1998 dataset is used to validate the 12 calibrated Bayesian recharge models. The modeled recharge is plotted versus the RORA recharge as shown in Figure 4. Results suggest a strong agreement between the Bayesian recharge models and the measured recharge at low recharge values and a weaker agreement at high recharge values as shown in Figure 4. Overall, the validated recharge models have a good performance. The validated Bayesian recharge models in the Cedar Basin have R^2 of 0.62–0.68, RMSE of 10.1–11.1 mm, d1 of 0.70–0.72, and PBIAS of 33–49% as shown in Table 4. On the other hand, the River Basin's recharge models have relatively better performance in the validation phase, with R^2 of 0.66–0.76, RMSE of 8.7–10.3 mm, d1 of 0.81–0.83, and PBIAS of 12–25% as shown in Table 4.

Even though there is no significant difference in performance among the 12 Bayesian recharge models, the EBL-based Bayesian models (i.e. m.fixed_g.EBL, m.random_g.EBL, and m.uniform_g.EBL) performed slightly better than the rest of the models in simulating recharge in the Cedar Basin during the validation phase, as shown in Table 4 and Figure 4(a). However, in the Rainy Basin, the hyper.UIP-based Bayesian models (i.e. m.fixed_g.hyper.UIP, m.random_g.hyper.UIP, and m.uniform_g.hyper.UIP) seem to have slightly outperformed the rest of the Bayesian recharge models in the validation phase, as shown in Table 4 and Figure 4(b). There is no significant difference in the Bayesian recharge models' RMSE values between Cedar and Rainy River Basins, as shown in Table 4. Therefore, the average deviation in Bayesian models recharge estimation seems to be uniform.

3.6. Relative importance of the climate models outside and inside the Bayesian framework

The performance indices of the 10 climate models outside the Bayesian framework are summarized in Table 2. The models with the strongest agreement with the observed recharge have the highest importance during averaging in the Bayesian framework. As found earlier, RCM3_gfdl has the best performance indices outside the Bayesian framework for the Cedar River Basin, and RCM3_cgcm3 is the best model for the Rainy River Basin. On the other hand, CRCM_ccsm performs the worst for both Cedar and Rainy Basins.

After averaging the climate models within the Bayesian frameworks, the relative importance of the models (within the framework) is evaluated based on their PIPs. The PIP values of the 10 RCMs are shown in Figure 5. The results suggest that MM5I_ccsm and RCM3_gfdl are the most important RCMs in the Cedar Basin since these RCMs have PIP values

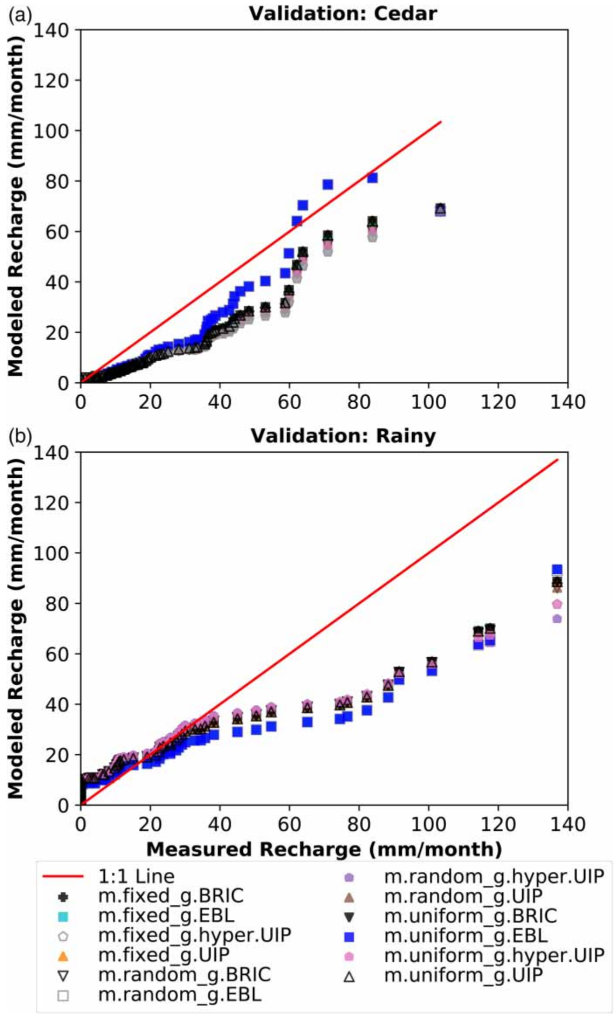

Figure 4 | Validated Bayesian recharge models for (a) Cedar and (b) Rainy River Basins using **1991 and 1998** data.

Table 4 | Performance of the validated Bayesian recharge models using **1991–1998** data

	Cedar				**Rainy**			
Bayesian models	R^2	**RMSE (mm)**	**d1 (–)**	**PBIAS (%)**	R^2 **(–)**	**RMSE (mm)**	**d1 (–)**	**PBIAS (%)**
m.fixed_g.BRIC	0.941	11.6	0.680	44.6	0.933	13.7	0.794	13.3
m.fixed_g.EBL	0.917	9.01	0.757	34.0	0.937	14.9	0.783	25.2
m.fixed_g.hyper.UIP	0.943	13.0	0.647	49.0	0.917	14.1	0.801	12.9
m.fixed_g.UIP	0.941	11.6	0.680	44.6	0.933	13.7	0.794	13.3
m.random_g.BRIC	0.942	12.4	0.661	47.0	0.930	13.8	0.795	13.2
m.random_g.EBL	0.917	9.01	0.757	33.9	0.937	14.9	0.784	25.3
m.random_g.hyper.UIP	0.942	12.4	0.661	47.0	0.903	14.6	0.800	14.0
m.random_g.UIP	0.942	12.4	0.661	47.0	0.930	13.8	0.795	13.2
m.uniform_g.BRIC	0.941	11.6	0.680	44.6	0.933	13.7	0.794	13.3
m.uniform_g.EBL	0.917	9.01	0.757	34.0	0.937	14.9	0.783	25.2
m.uniform_g.hyper.UIP	0.943	12.953	0.647	49.0	0.917	14.1	0.801	12.9
m.uniform_g.UIP	0.941	11.636	0.680	44.6	0.933	13.7	0.794	13.3

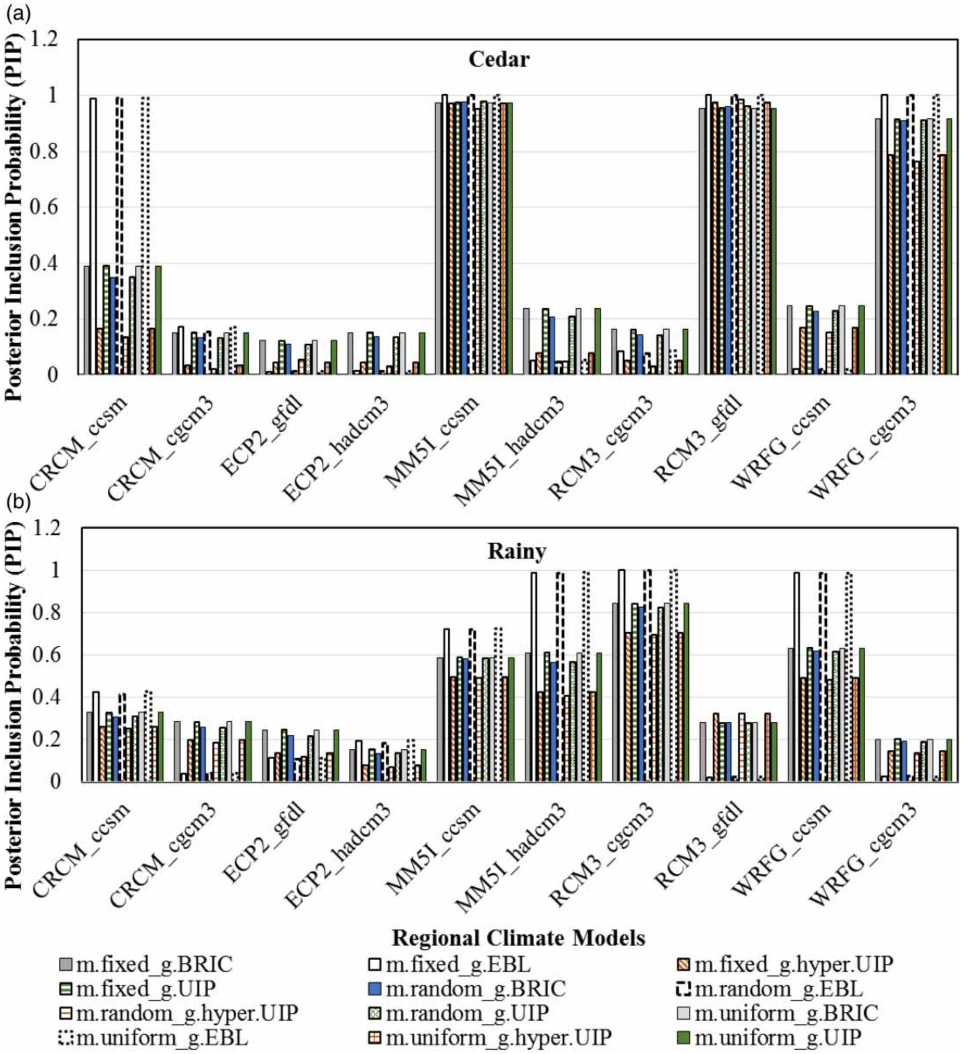

Figure 5 | PIP of **trained** Bayesian models for (a) Cedar and (b) Rainy Basins using (1968–1996) data.

of close to 1 in nearly all the 12 Bayesian recharge models, as shown in Figure 5(a). Note that even though the MM5I_ccsm does not have the strongest agreement with observed recharge outside the Bayesian framework, its performance indices values suggest a good performance based on R^2, RMSE, and d1. Beside the two best performing RCMs, WRFG_cgcm3 also has a good performance in the Cedar Basin with PIP values of 0.8–1 across the 12 Bayesian models, as shown in Figure 5(a). The remaining seven RCMs for the Cedar Basin have a weak performance with PIP values of less than 0.5, as shown in Figure 5(a). High variability of PIP values is also observed among the weak performing RCMs since the PIP values are in the range of 0–0.25 across the 12 Bayesian recharge models for each of the weak performing RCMs. CRCM_ccsm is also categorized as a weak performing model in the Cedar Basin since its PIP values from 7 of the 12 Bayesian models are less than 0.5. This is consistent with its performance indices values outside the Bayesian framework. However, the EBL-based Bayesian models (i.e. m.fixed_g.EBL, m.random_g.EBL, and m.uniform_g.EBL) assign PIP values of 1 to the weak performing CRCM_ccsm. This is due to the fact that EBL g-prior uses a local approach to compute g-prior based on the *F*-statistics of the individual regression models within the Bayesian regression model space of 2^{10} since there are 10 RCMs. The rest of the Bayesian models use a global approach in which a common g-value across all the regression models within the model space is applied.

In the Rainy River Basin, the best performing climate model is RCM3_cgcm3 since it has the PIP values of 0.75–1 across the 12 Bayesian recharge models, as shown in Figure 5(b). This is followed by MM5I_ccsm, MM5I_hadcm3, and WRFG_ccsm, which have PIP values of 0.5–0.7, as shown in Figure 5(b). The worst performing RCMs outside the Bayesian

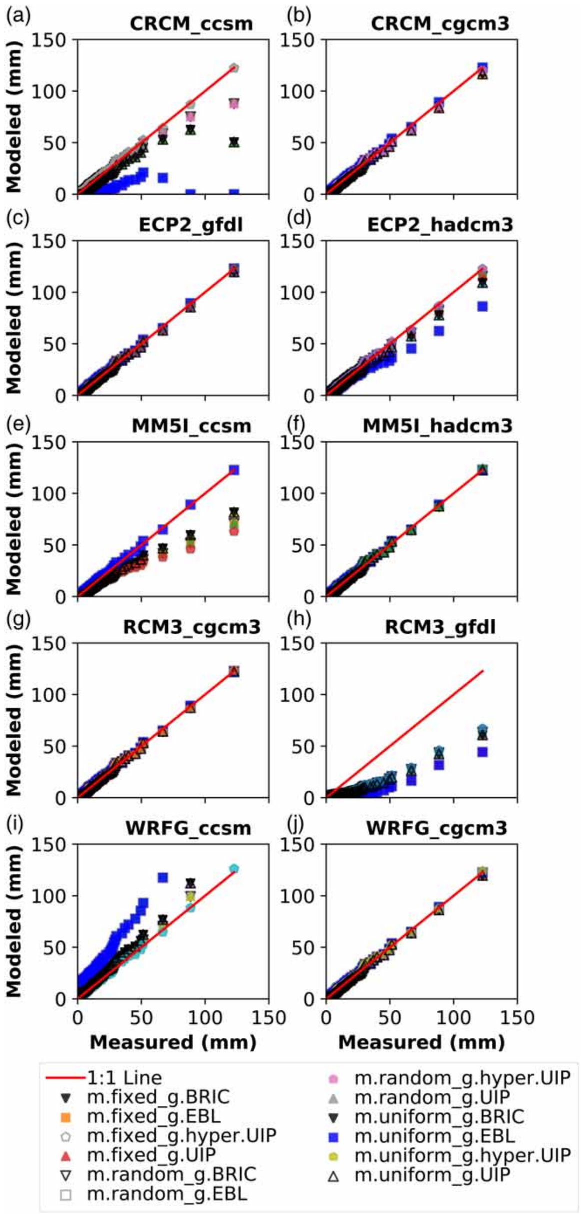

Figure 6 | Sensitivity of Bayesian recharge models to the RCMs in the Cedar River Basin.

framework, CRCM_ccsm, also has the smallest PIP values (PIP value of <0.5), as shown in Figure 5(b). Therefore, the PIP values of the poor performing RCMs seem to be consistent with the models performance outside the Bayesian framework. The remaining RCMs have weak performance with PIP values of <0.5, as shown in Figure 5(b). The EBL-based Bayesian models seem to exaggerate the PIP values in both best and worst performing RCMs in the Rainy Basin, as shown in Figure 5(b). As explained earlier, the slightly larger PIP values observed in the EBL-based Bayesian models are attributed to the fact that the g-value of the EBL prior is computed locally for each regression model, unlike the non-Empirical Bayes g-Local (non-EBL)-based Bayesian models. The recharge estimations are relatively more uncertain in the Rainy River Basin than in the Cedar Basin. This is observed in the relatively high coefficient of variation of recharge prediction in the Rainy Basin of 1.42 than in the Cedar Basin of 1.23. The relatively high recharge prediction variability among the climate models explains the observed variability of PIP values across the 12 Bayesian recharge models for almost every climate model, as shown in Figure 5(b). Even though RCMs' PIP values are highly variable in the Rainy Basin, the Bayesian models still give a clear direction of the recharge prediction performance strength of the RCMs within the Bayesian framework.

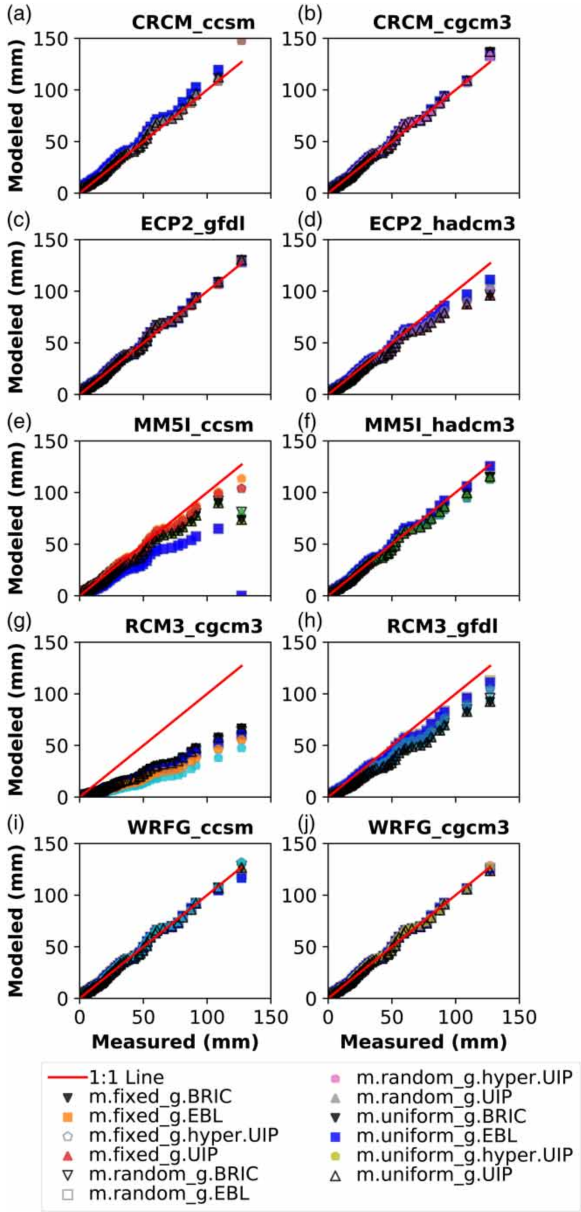

Figure 7 | Sensitivity of Bayesian recharge models to the RCMs in the Rainy River Basin.

3.7. Sensitivity of Bayesian recharge models to the climate models

Based on a leave-one-out technique, recharge prediction of the Bayesian recharge models was computed by leaving out one of the 10 RCMs. The sensitivity of the Bayesian recharge models was expressed as both scatter plots (Figures 6 and 7) and mean percent change (Figures 8 and 9) of Bayesian recharge prediction between the Bayesian recharge prediction and the leave-one-out-based Bayesian recharge prediction. The computation of the mean percent change of recharge is shown in equation A4 in the Appendix.

Figures 6 and 7 are plots of modeled recharge with one climate model left out versus the observed recharge for Cedar and Rainy River Basins, respectively. In the Cedar River Basin, the Bayesian recharge models are most sensitive to WRFG_ccsm since the largest deviation of the modeled recharge from the observed recharge is observed when WRFG_ccsm is left out of the calibrated Bayesian recharge models, as shown in Figures 6(i) and 8(b). It is also observed that the changes in recharge when WRFG_ccsm is left out the Bayesian models lead to the overestimation of recharge in the Cedar River Basin. The worst performing model inside and outside the Bayesian framework, CRCM_ccsm, seems to produce low-to-moderate sensitivity, as

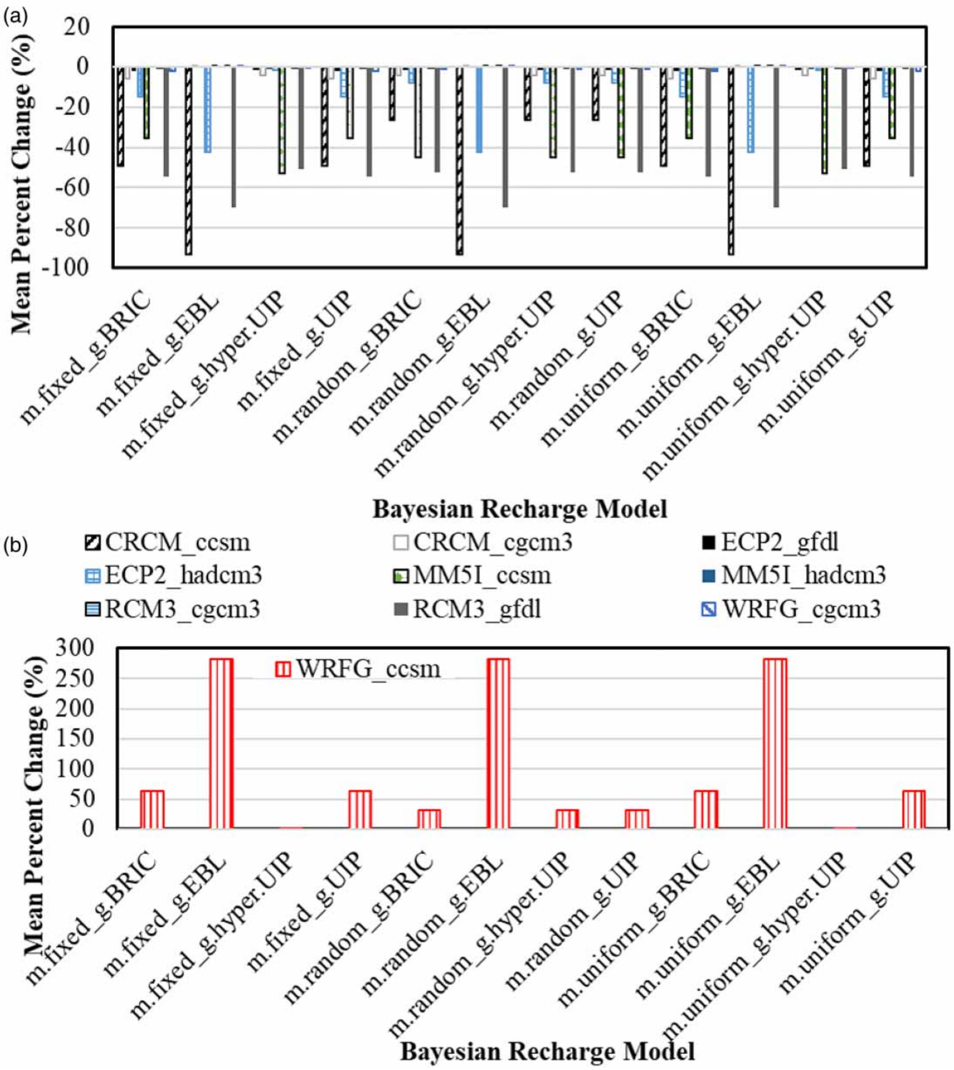

Figure 8 | Mean percent change (%) of Bayesian recharge models due to leave-one-out of the RCMs in the Cedar River Basin. (a) RCMs that produce low-to-moderate sensitivity and (b) RCM that produces the highest sensitivity.

shown in Figures 8 and 9 when left out of the Bayesian recharge model. However, this model is still important within the Bayesian framework since leaving CRCM_ccsm out of the calibrated Bayesian models causes underestimation of recharge in the Cedar Basin, especially at high recharge values, as shown in Figure 6(a). The only climate model that results in overestimation in recharge in the Cedar Basin, when the model is left out, is WRFG_ccsm, as shown in Figures 6(i) and 8(b). Note that the overestimation of recharge is mostly significant at high recharge values when the WRFG_ccsm model is left out. An insignificant change in recharge occurs when CRCM_cgcm3, ECP2_gfdl, ECP2_hadcm3, MM5I_hadcm3, or RCM3_cgcm3 is left out, as shown in Figure 6.

Like the Cedar River Basin, leaving out ECP2_gfdl does not cause any significant change in recharge prediction by Bayesian recharge models in the Rainy Basin, as shown in Figure 7(c). The largest underestimation in recharge in the Rainy River Basin occurs when RCM3_cgcm3 is left out, as shown in Figure 7(g). However, this underestimation does not deviate much from the original prediction (when none of the RCMs is left out). Leaving out CRCM_ccsm produces the largest mean percent change in recharge prediction, in the Rainy Basin, as shown in Figure 9(b). Therefore, in the Rainy Basin, Bayesian recharge models are most sensitive to CRCM_ccsm. A slight overestimation and underestimation of recharge, in the lower and upper recharge values, respectively, is observed when ECP2_hadcm3 is left out of the Bayesian recharge models, as shown in Figure 7(d). This is also supported by the relatively moderate mean percent change (less than −20%) as shown in Figure 9(a). No significant change in recharge is observed when WRFG_cgcm3 is left out, as shown in Figures 7(j) and 9(a).

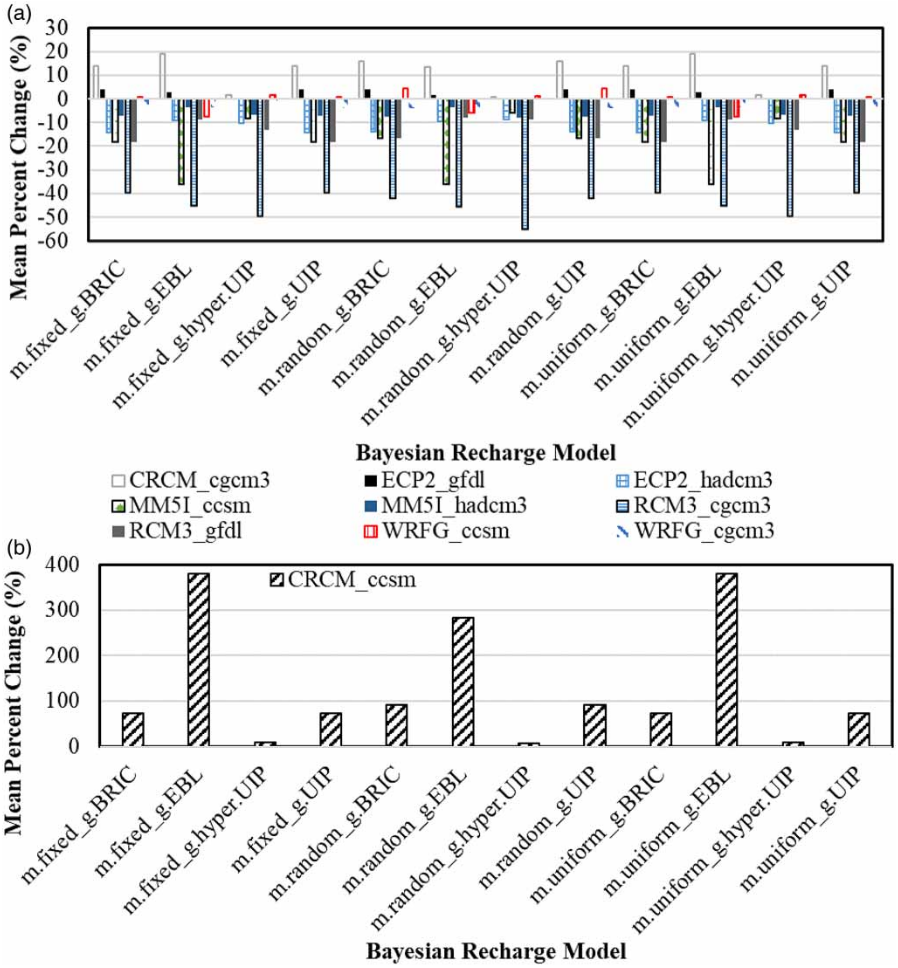

Figure 9 | Mean percent change (%) of Bayesian recharge models due to leave-one-out of the RCMs in the Rainy River Basin. (a) RCMs that produce low-to-moderate sensitivity and (b) RCM that produces the highest sensitivity.

In the Cedar River Basin, the Bayesian recharge models are most sensitive to RCM3_gfdl because RCM3_gfdl has the largest Bayesian regression coefficient value of an average of 0.56 mm/month. Therefore, removing RCM3_gfdl from the calibrated Bayesian recharge models leads to the greatest change in the modeled recharge. Unlike the RCM3_gfdl, leaving out WRFG_ccsm only produces strong sensitivity in EBL-based recharge models depicted by the relatively large mean percent change of recharge. On the other hand, ECP2_gfdl has the smallest Bayesian regression coefficient value of 0.01 mm/month and thus leads to the smallest deviation when being left out of the calibrated Bayesian recharge models.

Within the calibrated Bayesian regression model, RCM3_cgcm3 has the largest regression coefficient – an average of 0.55 mm/month – across all Bayesian recharge models. Therefore, removing RCM3_cgcm3 from the Bayesian recharge models causes the greatest change in the predicted recharge in the Rainy Basin. This is followed closely by CRCM_ccsm which, unlike the RCM3_cgcm3, only produces high sensitivity in the EBL-based recharge models of the Rainy Basin. Conversely, since ECP2_hadcm3 has the smallest Bayesian regression coefficient of −0.01 mm/month and leaving it out leads to the smallest change in the Bayesian modeled recharge in the Rainy River Basin. Except for WRFG_ccsm and CRCM_ccsm models, leaving out the climate models from the Bayesian recharge models mainly causes underestimation in the recharge in the Rainy River Basin. In the Cedar Basin, leaving out WRFG_ccsm from the Bayesian models could cause overestimation in the recharge because WRFG_ccsm has a negative effect due to the relatively larger negative PBIAS of −6.79%. Moreover, WRFG_ccsm has the largest negative regression coefficient of −0.22 mm/month. Therefore, removing WRFG_ccsm removes the negative effect on the Bayesian recharge models and thus overestimation of the recharge in the Rainy River Basin. The negative bias in recharge prediction can be explained by the tendency of WRFG_ccsm's LSM (NOAH) to overestimate surface runoff due to the assumed frozen top soil layer in model physics (Niu *et al.* 2011), which makes it more impermeable than its counterpart LSMs.

3.8. Implication of averaging climate model-based recharge in the Bayesian framework

Groundwater recharge replenishes aquifer storage, thus sustaining demand from the ever-increasing global population, alleviating potential environmental impact like land subsidence, and dampening of draining of nearby rivers that are hydraulically connected to the aquifer. Traditional methods of estimating recharge (e.g. lysimeters and isotope techniques) are often expensive and not sustainable in long-term recharge monitoring, especially in the face of climate change. Therefore, climate models come in handy in monitoring groundwater recharge, not only in the current climate scenario, but also in future climate scenarios. However, individual climate models often produce different recharge predictions. As a result, most recharge studies often use model selection to select climate model that provides the closest recharge prediction. Model selection leads to the elimination of the climate models whose predictions are relatively over- and underestimations of the actual recharge. Climate models are costly to develop with respect to time, knowledge, and labor. Eliminating climate models that do not give predictions that are close to the measured recharge is merely a waste of resources. In this study, out of the 10 climate models, ECP2_hadcm3 and RCM3_cgcm3 would be selected since they provide recharge that is closest to the observations. However, this will mean that the rest of the models are discarded. To harness the potential relative importance of all the 10 models, we averaged their recharge predictions in a Bayesian framework. However, because Bayesian priors have been found to impact posterior and thus predictions (Kavetski *et al.* 2006; Bastola *et al.* 2011), it is important to identify the prior combinations (MPrior and g-prior) that would best average recharge predictions from climate models in a Bayesian framework. Averaging the recharge predictions in the Bayesian framework also allows us to determine the relative contribution of the climate models to the Bayesian recharge model. As a result, we not only used all the climate models in producing superior recharge predictions by identifying the most suitable Bayesian prior combinations for recharge but also provided opportunity for improving future versions of these climate models based on their relative contribution to the Bayesian recharge model.

4. CONCLUSIONS

Our study suggests that non-EBL-based Bayesian frameworks are suitable for averaging recharge from climate models within a Bayesian framework. The performance of the climate models (in modellng recharge) inside the non-EBL-based Bayesian framework is consistent with their corresponding performance outside the Bayesian framework. Inconsistency between climate models' performance inside and outside the Bayesian framework arising from the EBL-based Bayesian frameworks can be attributed to the fact that their g-parameter is regression model-specific, whereas a constant g-parameter is used (for all the 1,024 regression models) in the non-EBL-based priors. These observations are consistent with previous studies (e.g. Achieng & Zhu 2019). However, since the previous study was on different regimes of streamflow, EBL priors was found to be just as suitable for estimating low flows as the non-EBL-based priors (Achieng & Zhu 2019). Note that the basin size does not seem to affect the choice of Bayesian framework when estimating the recharge in the Bayesian framework – unlike the basin size-effect on the choice of Bayesian framework for streamflow in the previous study (Achieng & Zhu 2019). This could be attributed to the fact that, unlike streamflow which is a surface hydrological process, recharge is a process that occurs much slower than streamflow and the recharge is relatively less affected than the streamflow by many hydro-topo-climatological factors. Therefore, the recharge is much stable and thus even though there was uncertainty in the climate model-based recharge prediction, all the climate models underestimated the recharge. Either way, the climate models' uncertainty is evident in both streamflow and recharge. This can be contributed to the different model physics and assumptions that are applied on the LSMs of the respective climate models and the different spatial resolution GCMs that provide the input to the corresponding RCMs. The RORA method has been found to provide reliable recharge values. However, previous studies have also found that the RORA method has some uncertainty due to many assumptions that the method has been developed under (Delin *et al.* 2007; Lorenz & Delin 2007). Future studies should focus on characterizing the uncertainties from the 'reference recharge' and the climate models themselves. The relative strength of climate models at predicting groundwater recharge can be harnessed by averaging their predictions within a Bayesian framework. However, it is important to use the Bayesian framework with priors that result into PIP consistent with the performance of the respective climate models outside the Bayesian framework because of difference in Bayesian frameworks with respect to prior probability of both Bayesian regression models and regression coefficients. In this study, we comprehensively evaluate the performance of groundwater recharge predictions from 12 different Bayesian frameworks. EBL-based prior is found to overestimate PIP values of the poor performing climate models. Non-EBL-based priors result in PIP values that are consistent with the performance of the climate models outside the Bayesian framework in the two case-study basins. Therefore, we recommend the use of non-EBL prior-based Bayesian frameworks when averaging recharge predictions from climate models within the Bayesian framework. Other conclusions can be summarized as follows:

- The choice of prior affects the suitability of Bayesian framework for averaging recharge from multiple climate models in a Bayesian framework.
- Averaging recharge predictions from climate models within the Bayesian framework results in better recharge prediction than that from any of the individual climate models.
- Nearly all the NARCCAP RCMs underestimate groundwater recharge.

ACKNOWLEDGEMENTS

This research was supported by the Bishop, Floyd & Wilma Endowment and Paul A. Rechard Fellowship at the University of Wyoming, USA.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

REFERENCES

Achieng, K. O. & Zhu, J. 2019 Application of Bayesian framework for evaluation of streamflow simulations using multiple climate models. *Journal of Hydrology* **574**, 1110–1128. https://doi.org/10.1016/j.jhydrol.2019.05.018.

Ajami, N. K., Duan, Q. & Sorooshian, S. 2007 An integrated hydrologic Bayesian multimodel combination framework: confronting input, parameter, and model structural uncertainty in hydrologic prediction. *Water Resources Research* **43** (1). https://doi.org/10.1029/2005WR004745.

Allen, D. M., Cannon, A. J., Toews, M. W. & Scibek, J. 2010 Variability in simulated recharge using different GCMs. *Water Resources Research* **46** (10). https://doi.org/10.1029/2009WR008932.

Arellano, L. N., Good, S. P., Sánchez-Murillo, R., Jarvis, W. T., Noone, D. C. & Finkenbiner, C. E. 2020 Bayesian estimates of the mean recharge elevations of water sources in the Central America region using stable water isotopes. *Journal of Hydrology: Regional Studies* **32**, 100739. https://doi.org/10.1016/j.ejrh.2020.100739.

Arnold, J. G. & Allen, P. M. 1999 Automated methods for estimating baseflow and ground water recharge from streamflow records. *Journal of the American Water Resources Association* **35** (2). https://doi.org/10.1111/j.1752-1688.1999.tb03599.x.

Bastola, S., Murphy, C. & Sweeney, J. 2011 The role of hydrological modelling uncertainties in climate change impact assessments of Irish river catchments. *Advances in Water Resources* **34** (5), 562–576. https://doi.org/10.1016/J.ADVWATRES.2011.01.008.

Berger, K. 2000 Validation of the hydrologic evaluation of landfill performance (HELP) model for simulating the water balance of cover systems. *Environmental Geology* **39**.

Berger, K. 2015 On the current state of the hydrologic evaluation of landfill performance (HELP) model. *Waste Management* **38**, 201–209. https://doi.org/10.1016/J.WASMAN.2015.01.013.

Brunner, P., Bauer, P., Eugster, M. & Kinzelbach, W. 2004 Using remote sensing to regionalize local precipitation recharge rates obtained from the Chloride Method. *Journal of Hydrology* **294** (4), 241–250. https://doi.org/10.1016/J.JHYDROL.2004.02.023.

Busenberg, E. & Plummer, L. N. 1992 Use of chlorofluorocarbons (CCl3F and CCl2F2) as hydrologic tracers and age-dating tools: the alluvium and terrace system of central Oklahoma. *Water Resources Research* **28** (9), 2257–2283. https://doi.org/10.1029/92WR01263.

Caya, D. & Laprise, R. 1999 A semi-implicit semi-Lagrangian regional climate model: the Canadian RCM. *Monthly Weather Review* **127** (3), 341–362. https://doi.org/10.1175/1520-0493(1999)127&lt;0341:ASISLR > 2.0.CO;2.

Collins, W. D., Bitz, C. M., Blackmon, M. L., Bonan, G. B., Bretherton, C. S., Carton, J. A. & … Smith, R. D. 2006 The Community Climate System Model Version 3 (CCSM3). *Journal of Climate* **19** (11), 2122–2143. https://doi.org/10.1175/JCLI3761.1.

Crosbie, R. S., McCallum, J. L., Walker, G. R. & Chiew, F. H. S. 2010 Modelling climate-change impacts on groundwater recharge in the Murray-Darling Basin, Australia. *Hydrogeology Journal* **18** (7), 1639–1656. https://doi.org/10.1007/s10040-010-0625-x.

Crosbie, R. S., Pickett, T., Mpelasoka, F. S., Hodgson, G., Charles, S. P. & Barron, O. V. 2013 An assessment of the climate change impacts on groundwater recharge at a continental scale using a probabilistic approach with an ensemble of GCMs. *Climatic Change* **117** (1–2), 41–53. https://doi.org/10.1007/s10584-012-0558-6.

Cui, W. & George, E. I. 2004 *Empirical Bayes vs. Fully Bayes Variable Selection*. Available from: http://www-stat.wharton.upenn.edu/~edgeorge/Research_papers/CGrevision12-06.pdf.

Daniel, J. F. 1976 Estimating groundwater evapotranspiration from streamflow records. *Water Resources Research* **12** (3), 360–364. https://doi.org/10.1029/WR012i003p00360.

Dawson, C. W., Harpham, C., Wilby, R. L. & Chen, Y. 2002 Evaluation of artificial neural network techniques for flow forecasting in the River Yangtze, China. *European Geosciences Union* **6**. Available from: https://hal.archives-ouvertes.fr/hal-00304714

Delin, G. N., Healy, R. W., Lorenz, D. L. & Nimmo, J. R. 2007 Comparison of local- to regional-scale estimates of ground-water recharge in Minnesota, USA. *Journal of Hydrology* **334** (1–2), 231–249. https://doi.org/10.1016/j.jhydrol.2006.10.010.

Delworth, T. L., Broccoli, A. J., Rosati, A., Stouffer, R. J., Balaji, V., Beesley, J. A. & … Zhang, R. 2006 GFDL's CM2 global coupled climate models. Part I: formulation and simulation characteristics. *Journal of Climate* **19** (5), 643–674. https://doi.org/10.1175/JCLI3629.1.

Dickinson, R. E. & Henderson-Sellers, A. 1988 Modelling tropical deforestation: a study of GCM land-surface parametrizations. *Quarterly Journal of the Royal Meteorological Society* **114** (480), 439–462. https://doi.org/10.1002/qj.49711448009.

Dickinson, R. E., Henderson-Sellers, A., Kennedy, J. & Wilson, F. 1986 *Biosphere-Atmosphere Transfer Scheme (BATS) for the NCAR Community Climate Model*. https://doi.org/10.5065/D6668B58.

Duan, Q. & Phillips, T. J. 2010 Bayesian estimation of local signal and noise in multimodel simulations of climate change. *Journal of Geophysical Research* **115** (D18), D18123. https://doi.org/10.1029/2009JD013654.

Ek, M. B., Mitchell, K. E., Lin, Y., Rogers, E., Grunmann, P., Koren, V. & … Tarpley, J. D. 2003 Implementation of NOAH land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. *Journal of Geophysical Research: Atmospheres* **108** (D22). https://doi.org/10.1029/2002JD003296.

Fernández, C., Ley, E. & Steel, M. F. J. 2001a Benchmark priors for Bayesian model averaging. *Journal of Econometrics* **100** (2), 381–427. https://doi.org/10.1016/S0304-4076(00)00076-2.

Fernández, C., Ley, E. & Steel, M. F. J. 2001b Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics* **16** (5), 563–576. https://doi.org/10.1002/jae.623.

Fernández, C., Ley, E. & Steel, M. F. J. 2002 Bayesian modelling of catch in a north-west Atlantic fishery. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **51** (3), 257–280. https://doi.org/10.1111/1467-9876.00268.

Flato, G. M., Boer, G. J., Lee, W. G., McFarlane, N. A., Ramsden, D., Reader, M. C. & Weaver, A. J. 2000 The Canadian centre for climate modelling and analysis global coupled model and its climate. *Climate Dynamics* **16** (6), 451–467. https://doi.org/10.1007/s003820050339.

George, E. I. & Foster, D. P. 2000 Calibration and empirical Bayes variable selection. *Biometrika* **87** (4), 731–747.

Grell, G., Dudhia, J. & Stauffer, D. 1994 *A Description of the Fifth-Generation Penn State/NCAR Mesoscale Model (MM5)*. https://doi.org/10.5065/D60Z716B.

Gupta, H. V., Sorooshian, S. & Yapo, P. O. 1998 Toward improved calibration of hydrologic models: multiple and noncommensurable measures of information. *Water Resources Research* **34** (4), 751–763. https://doi.org/10.1029/97WR03495.

Ju, Y. J., Massoudieh, A., Green, C. T., Lee, K. K. & Kaown, D. 2021 Complexity of groundwater age mixing near a seawater intrusion zone based on multiple tracers and Bayesian inference. *Science of the Total Environment* **753**, 141994. https://doi.org/10.1016/j.scitotenv.2020.141994.

Juang, H.-M. H., Hong, S.-Y., Kanamitsu, M., Juang, H.-M. H., Hong, S.-Y. & Kanamitsu, M. 1997 The NCEP regional spectral model: an update. *Bulletin of the American Meteorological Society* **78** (10), 2125–2143. https://doi.org/10.1175/1520-0477(1997)078&lt;2125:TNRSMA > 2.0.CO;2.

Jyrkama, M. I. & Sykes, J. F. 2007 The impact of climate change on spatially varying groundwater recharge in the grand river watershed (Ontario). *Journal of Hydrology* **338** (3–4), 237–250. https://doi.org/10.1016/j.jhydrol.2007.02.036.

Kass, R. E. & Wasserman, L. 1995 A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association* **90** (431), 773–795. https://doi.org/10.2307/2291327.

Kavetski, D., Kuczera, G. & Franks, S. W. 2006 Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. *Water Resources Research* **42** (3). https://doi.org/10.1029/2005WR004368.

Legates, D. R. & McCabe, G. J. 1999 Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model validation. *Water Resources Research* **35** (1), 233–241. https://doi.org/10.1029/1998WR900018.

Ley, E. & Steel, M. F. J. 2008 On the effect of prior assumptions in Bayesian model averaging with applications to growth regression, **7**. Available from: http://www.warwick.ac.uk/go/msteel/.

Ley, E. & Steel, M. F. J. 2012 *Mixtures of g-Priors for Bayesian Model Averaging with Economic Applications*. Available from: https://mpra.ub.uni-muenchen.de/36817/.

Ley, E., Steel, M. F. J. J., Mark, A. & Steel, F. J. 2012 Mixtures of g-priors for Bayesian model averaging with economic applications. *Journal of Econometrics* **171** (2), 251–266. https://doi.org/10.1016/J.JECONOM.2012.06.009.

Liang, X., Wood, E. F., Lettenmaier, D. P., Lohmann, D., Boone, A., Chang, S. & … Zeng, Q.-C. 1998 The project for intercomparison of land-surface parameterization schemes (PILPS) phase-2(c) Red-Arkansas River basin experiment: 2. Spatial and temporal analysis of energy fluxes. *Global and Planetary Change* **19**, 137–159.

Liang, F., Paulo, R., Molina, G., Clyde, M. A. & Berger, J. O. 2008 Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association* **103** (481), 410–423. https://doi.org/10.1198/016214507000001337.

Lohmann, D., Lettenmaier, D. P., Liang, X., Wood, E. F., Boone, A., Chang, S. & … Zeng, Q.-C. 1998 The Project for Intercomparison of Land-surface Parameterization Schemes (PILPS) phase (2) Red–Arkansas River basin experiment: 3. Spatial and temporal analysis of water fluxes. *Global and Planetary Change* **19**, 161–179.

Long, D., Yang, W., Scanlon, B. R., Zhao, J., Liu, D., Burek, P. & … Wada, Y. 2020 South-to-North water diversion stabilizing Beijing's groundwater levels. *Nature Communications* **11** (1), 1–10. https://doi.org/10.1038/s41467-020-17428-6.

Lorenz, D. & Delin, G. N. 2007 A regression model to estimate regional ground water recharge. *Ground Water* **45** (2), 196–208. https://doi.org/10.1111/j.1745-6584.2006.00273.x.

Luo, L., Wood, E. F. & Pan, M. 2007 Bayesian merging of multiple climate model forecasts for seasonal hydrological predictions. *Journal of Geophysical Research: Atmospheres* **112** (D10). https://doi.org/10.1029/2006JD007655.

Ma, F., Ye, A., Deng, X., Zhou, Z., Liu, X., Duan, Q. & … Gong, W. 2016 Evaluating the skill of NMME seasonal precipitation ensemble predictions for 17 hydroclimatic regions in continental China. *International Journal of Climatology* **36** (1), 132–144. https://doi.org/10.1002/joc.4333.

Mearns, L. O., Arritt, R., Biner, S., Bukovsky, M. S., McGinnis, S., Sain, S. & … Snyder, M. 2012 The North American regional climate change assessment program dataset. *American Meteorological Society* **93** (9), 1337–1362. https://doi.org/10.1175/BAMS-D-11-00223.1.

Mearns, L. O., Sain, S., Leung, L. R., Bukovsky, M. S., McGinnis, S., Biner, S. & … Sloan, L. 2013 Climate change projections of the North American Regional Climate Change Assessment Program (NARCCAP). *Climatic Change* **120** (4), 965–975. https://doi.org/10.1007/s10584-013-0831-3.

Mehrnegar, N., Jones, O., Singer, M. B., Schumacher, M., Jagdhuber, T., Scanlon, B. R. & … Forootan, E. 2021 Exploring groundwater and soil water storage changes across the CONUS at 12.5 km resolution by a Bayesian integration of GRACE data into W3RA. *Science of the Total Environment* **758**, 143579. https://doi.org/10.1016/j.scitotenv.2020.143579.

Meixner, T., Manning, A. H., Stonestrom, D. A., Allen, D. M., Ajami, H., Blasch, K. W. & … Walvoord, M. A. 2016 Implications of projected climate change for groundwater recharge in the western United States. *Journal of Hydrology* **534**, 124–138. https://doi.org/10.1016/j.jhydrol.2015.12.027.

Molina, J. L., Bromley, J., García-Aróstegui, J. L., Sullivan, C. & Benavente, J. 2010 Integrated water resources management of overexploited hydrogeological systems using Object-Oriented Bayesian Networks. *Environmental Modelling and Software* **25** (4), 383–397. https://doi.org/10.1016/j.envsoft.2009.10.007.

Molina, J. L., Pulido-Velázquez, D., García-Aróstegui, J. L. & Pulido-Velázquez, M. 2013 Dynamic Bayesian networks as a decision support tool for assessing climate change impacts on highly stressed groundwater systems. *Journal of Hydrology* **479**, 113–129. https://doi.org/10.1016/j.jhydrol.2012.11.038.

Music, B. & Caya, D. 2007 Evaluation of the hydrological cycle over the Mississippi River Basin as simulated by the Canadian regional climate model (CRCM). *Journal of Hydrometeorology* **8** (5), 969–988. https://doi.org/10.1175/JHM627.1.

Mustafa, S. M. T., Nossent, J., Ghysels, G. & Huysmans, M. 2020 Integrated Bayesian multi-model approach to quantify input, parameter and conceptual model structure uncertainty in groundwater modeling. *Environmental Modelling and Software* **126**, 104654. https://doi.org/10.1016/j.envsoft.2020.104654.

NARCCAP. 2007 *North American Regional Climate Change Assessment Program (NARCCAP): Data Tables*. Available from: http://www.narccap.ucar.edu/data/data-tables.html (retrieved 12 April 2019).

Niraula, R., Meixner, T., Dominguez, F., Bhattarai, N., Rodell, M., Ajami, H. & … Castro, C. 2017 How might recharge change under projected climate change in the Western U.S.? *Geophysical Research Letters* **44** (20), 10,407–10,418. https://doi.org/10.1002/2017GL075421.

Niu, G.-Y., Yang, Z.-L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M. & … Xia, Y. 2011 The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *Journal of Geophysical Research* **116** (D12), D12109. https://doi.org/10.1029/2010JD015139.

Pal, J. S., Giorgi, F., Bi, X., Elguindi, N., Solmon, F., Gao, X. & … Steiner, A. L. 2007 *Regional Climate Modeling for the Developing World. The ICTP RegCM3 and RegCNET*. Available from: http://journals.ametsoc.org/doi/pdf/10.1175/BAMS-88-9-1395.

Pope, V. D., Gallani, M. L., Rowntree, P. R. & Stratton, R. A. 2000 The impact of new physical parametrizations in the Hadley Centre climate model: HadAM3. *Climate Dynamics* **16** (2–3), 123–146. https://doi.org/10.1007/s003820050009.

Risser, D. W., Gburek, W. J. & Folmar, G. J. 2005 *Ground-Water Resources Program Comparison of Methods for Estimating Ground-Water Recharge and Base Flow at a Small Watershed Underlain by Fractured Bedrock in the Eastern United States*. Available from: https://pubs.usgs.gov/sir/2005/5038/pdf/sir2005-5038.pdf.

Rorabaugh, M. I. 1964 Estimating changes in bank storage and ground-water contribution to streamflow. In: *International Association of Scientific Hydrology*, pp. 432–441.

Rorabaugh, M. & Simons, W. D. 1966 *Exploration of Methods of Relating Ground Water to Surface Water, Columbia River Basin-Second Phase. Open-File Report*. U.S. Dept. of the Interior. Available from: https://pubs.er.usgs.gov/publication/ofr66117.

Rosegrant, M. W., Ringler, C. & Zhu, T. 2009 Water for agriculture: maintaining food security under growing scarcity. *Annual Review of Environment and Resources* **34** (1), 205–222. https://doi.org/10.1146/annurev.environ.030308.090351.

Rutledge, A. T. 2000 *Considerations for Use of the Rora Program to Estimate Ground-Water Recharge from Streamflow Records*, Vol. 44.

Rutledge, A. T. 2004 *Use of RORA for Complex Ground-Water Flow Conditions*. U.S. Geological Survey Water-Resources Investigations Report 2003-4304. Available from: https://pubs.usgs.gov/wri/wri034304/pdf/wri03-4304_14_jan.pdf.

Rutledge, A. T. 2007 Update on the use of the RORA program for recharge estimation. *Ground Water* **45** (3), 374–382. https://doi.org/10.1111/j.1745-6584.2006.00294.x.

Scibek, J. & Allen, D. M. 2006 Comparing modelled responses of two high-permeability, unconfined aquifers to predicted climate change. *Global and Planetary Change* **50** (1–2), 50–62. https://doi.org/10.1016/J.GLOPLACHA.2005.10.002.

Siebert, S., Burke, J., Faures, J. M., Frenken, K., Hoogeveen, J., Döll, P. & Portmann, F. T. 2010 Groundwater use for irrigation – a global inventory. *Hydrology and Earth System Sciences* **14** (10), 1863–1880. https://doi.org/10.5194/hess-14-1863-2010.

Singh, N. K., Bhattacharya, R. & Borrok, D. M. 2020 A Bayesian framework to unravel food, groundwater, and climate linkages: a case study from Louisiana. *PLoS ONE* **15** (7), e0236757. https://doi.org/10.1371/journal.pone.0236757.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Wang, W. & Powers, J. G. 2005 *A Description of the Advanced Research WRF Version 2*. NCAR Tech Notes-468 + STR. Available from: http://www2.mmm.ucar.edu/wrf/users/docs/arw_v2.pdf.

Sloughter, M. J., Raftery, A. E., Gneiting, T. & Fraley, C. 2007 Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *American Meteorological Society* **135**, 3209–3220. https://doi.org/10.1175/MWR3441.1.

Steel, M. F. J. 2013 *Bayesian Model Averaging and Forecasting*. Available from: http://www.warwick.ac.uk/go/msteel/.

Verseghy, D. L. CLASS – a Canadian Land Surface Scheme for GCMS. I. Soil model. *International Journal of Climatology* **11** (2), 111–133. https://doi.org/10.1002/joc.3370110202.

Verseghy, D. L., McFarlane, N. A. & Lazare, M. 1993 CLASS – a Canadian Land Surface Scheme for GCMS, II. Vegetation model and coupled runs. *International Journal of Climatology* **13** (4), 347–370. https://doi.org/10.1002/joc.3370130402.

Vrugt, J. A., ter Braak, C. J. F., Clark, M. P., Hyman, J. M. & Robinson, B. A. 2008 Treatment of input uncertainty in hydrologic modeling: doing hydrology backward with Markov chain Monte Carlo simulation. *Water Resources Research* **44** (12). https://doi.org/10.1029/2007WR006720.

Wang, H., Lu, K., Zhao, Y., Zhang, J., Hua, J. & Lin, X. 2020 Multi-model ensemble simulated non-point source pollution based on Bayesian model averaging method and model uncertainty analysis. *Environmental Science and Pollution Research* **27** (35), 44482–44493. https://doi.org/10.1007/s11356-020-10336-8.

Willmott, C. J. 1982 Some comments on the evaluation of model performance. *Bulletin of the American Meteorological Society* **63** (11), 1309–1313. https://doi.org/10.1175/1520-0477(1982)063&lt;1309:SCOTEO > 2.0.CO;2.

Willmott, C. J., Ackleson, S. G., Davis, R. E., Feddema, J. J., Klink, K. M., Legates, D. R. & … Rowe, C. M. 1985 Statistics for the evaluation and comparison of models. *Journal of Geophysical Research* **90** (C5), 8995. https://doi.org/10.1029/JC090iC05p08995.

Woyshner, M. R. & Yanful, E. K. 1995 Modelling and field measurements of water percolation through an experimental soil cover on mine tailings. *Canadian Geotechnical Journal* **32**, 601–609.

Xie, Y., Cook, P. G., Simmons, C. T., Partington, D., Crosbie, R. & Batelaan, O. 2018 Uncertainty of groundwater recharge estimated from a water and energy balance model. *Journal of Hydrology* **561**, 1081–1093. https://doi.org/10.1016/J.JHYDROL.2017.08.010.

Yang, Z.-L. & Dickinson, R. E. 1996 Description of the biosphere-atmosphere transfer scheme (BATS) for the soil moisture workshop and evaluation of its performance. *Global and Planetary Change* **13** (1–4), 117–134. https://doi.org/10.1016/0921-8181(95)00041-0.

Yang, Z.-L., Niu, G.-Y., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M. & … Xia, Y. 2011 The community Noah land surface model with multiparameterization options (Noah-MP): 2. Evaluation over global river basins. *Journal of Geophysical Research* **116** (D12), D12110. https://doi.org/10.1029/2010JD015140.

Yin, J., Medellín-Azuara, J., Escriva-Bou, A. & Liu, Z. 2021 Bayesian machine learning ensemble approach to quantify model uncertainty in predicting groundwater storage change. *Science of The Total Environment* **769**, 144715. https://doi.org/10.1016/j.scitotenv.2020.144715.

Zeugner, S. 2011 *Bayesian Model Averaging with BMS*. Available from: https://cran.r-project.org/web/packages/BMS/vignettes/bms.pdf.

Zeugner, S. & Feldkircher, M. 2015 Bayesian model averaging employing fixed and flexible priors: the BMS package for R. *Journal of Statistical Software* **68** (4), 1–37. https://doi.org/10.18637/jss.v068.i04.

Zhang, L. & Dawes, W. 1998 *WAVES – An Integrated Energy and Water Balance Model*. Technical Report No. 31/98, CSIRO Land and Water, Canberra, Australia. Available from: https://publications.csiro.au/rpr/download?pid=procite:15e98eaa-36b5-47ef-9d0c-ab932796385e&dsid=DS1.

Zhang, L., Yang, X., Zhang, L. & Yang, X. 2018 Applying a multi-model ensemble method for long-term runoff prediction under climate change scenarios for the Yellow River Basin, China. *Water* **10** (3), 301. https://doi.org/10.3390/w10030301.

First received 19 December 2020; accepted in revised form 25 August 2021. Available online 7 September 2021