AGU PUBLICATIONS

JGR

Journal of Geophysical Research: Atmospheres

RESEARCH ARTICLE
10.1002/2015JD023401

Key Points:

- Evaluation of global impact models in reproducing runoff over central U.S.
- Overall, GIMs can reproduce trends in high, medium, and low flows well
- Skill of GIMs differs (water only/energy balance) in capturing runoff seasonality

Supporting Information:

- Figures S1–S33 and Table S1

Correspondence to:
I. Giuntoli,
ignazio.g@gmail.com

Citation:
Giuntoli, I., G. Villarini, C. Prudhomme, I. Mallakpour, and D. M. Hannah (2015), Evaluation of global impact models' ability to reproduce runoff characteristics over the central United States, *J. Geophys. Res. Atmos.*, *120*, 9138–9159, doi:10.1002/2015JD023401.

Received 18 MAR 2015
Accepted 12 AUG 2015
Accepted article online 15 AUG 2015
Published online 19 SEP 2015

Evaluation of global impact models' ability to reproduce runoff characteristics over the central United States

Ignazio Giuntoli[1,2,3], Gabriele Villarini[3], Christel Prudhomme[2], Iman Mallakpour[3], and David M. Hannah[1]

[1]School of Geography, Earth, and Environmental Sciences, University of Birmingham, Birmingham, UK, [2]Centre for Ecology and Hydrology, Wallingford, UK, [3]IIHR-Hydroscience and Engineering, University of Iowa, Iowa City, Iowa, USA

Abstract The central United States experiences a wide array of hydrological extremes, with the 1993, 2008, 2013, and 2014 flooding events and the 1988 and 2012 droughts representing some of the most recent extremes, and is an area where water availability is critical for agricultural production. This study aims to evaluate the ability of a set of global impact models (GIMs) from the Water Model Intercomparison Project to reproduce the regional hydrology of the central United States for the period 1963–2001. Hydrological indices describing annual daily maximum, medium and minimum flow, and their timing are extracted from both modeled daily runoff data by nine GIMs and from observed daily streamflow measured at 252 river gauges. We compare trend patterns for these indices, and their ability to capture runoff volume differences for the 1988 drought and 1993 flood. In addition, we use a subset of 128 gauges and corresponding grid cells to perform a detailed evaluation of the models on a gauge-to-grid cell basis. Results indicate that these GIMs capture the overall trends in high, medium, and low flows well. However, the models differ from observations with respect to the timing of high and medium flows. More specifically, GIMs that only include water balance tend to be closer to the observations than GIMs that also include the energy balance. In general, as it would be expected, the performance of the GIMs is the best when describing medium flows, as opposed to the two ends of the runoff spectrum. With regards to low flows, some of the GIMs have considerably large pools of zeros or low values in their time series, undermining their ability in capturing low flow characteristics and weakening the ensemble's output. Overall, this study provides a valuable examination of the capability of GIMs to reproduce observed regional hydrology over a range of quantities for the central United States.

1. Introduction

Freshwaters play a vital role in our lives and that of the ecosystems. In addition to drinking and sanitation, water is needed for economic activities such as agriculture and industry and for power production. There is a growing consensus that an intensification of the hydrological cycle is occurring [e.g., *Held and Soden*, 2006; *Huntington*, 2006; *Stott et al*., 2010]. As a result, hydrological extremes are likely to become more frequent [e.g., *Christensen and Christensen*, 2003; *Milly et al*., 2002, 2005; *Mallakpour and Villarini*, 2015], potentially leading to disruptive impacts on economic activities and a large toll in terms of casualties and damage to infrastructures. In this context, a better understanding of the present and future hydrological processes is ever more crucial for anticipating and taking necessary mitigation and adaptation measures. A valuable contribution in this direction is provided by global impact models (GIMs), which allow simulation of the terrestrial water cycle at the global scale. Together with global circulation models (GCMs), GIMs represent the physical processes in the atmosphere and land surface and operate over relatively long time span (decades), at a coarse spatial resolution (typically 50–250 km), and time step from subdaily to monthly. Broadly speaking, GIMs focus on simulating the land surface whereas GCMs focus primarily on the atmosphere (although they generally include some sort of land surface scheme, usually less sophisticated than that of the GIMs). Regarding the water cycle, the two model families meet at the land surface/atmosphere interface, which represents the upper boundary for the GIMs and the lower boundary for the GCMs. Therefore, GCM climate outputs often provide the basis for impact studies, in which GIMs consider the interaction of the atmospheric and land surface component of the water cycle [e.g., *Mölders*, 2005].

GIMs can be subdivided into two broad categories, which differ in the land surface parameterizations: (i) the global hydrological models (GHMs) have the water budget and lateral transfer of water as the main interest, requiring a partitioning of precipitation into evapotranspiration, infiltration, interception, storage, and runoff to determine the water fluxes within the soil and the groundwater recharge, and (ii) the land surface models (LSMs) try additionally to close the energy budget and run at subdaily time steps. With the aim to describe the vertical exchanges of heat, water, and sometimes carbon in considerable details, LSMs need a partitioning for precipitation between the aforementioned processes to determine the partitioning of radiative forcing between soil heat flux and the turbulent fluxes of sensible and latent heat [e.g., *Mölders*, 2005].

In the recent past, the hydrological impact research community has realized that the uncertainty associated with the GIMs (including model parameterization and structure) could be large and should not be neglected [*Prudhomme and Davies*, 2008]. It has also been recognized that multimodel ensembles are much more robust tools to address the uncertainty associated with climate change impact than single models and hence should be used as much as possible in any climate change assessment work [e.g., *Hagemann et al*., 2013]. At the local/catchment scale, this is achieved through building hydrological catchment model ensembles [e.g., *Smith et al*., 2012] from a wide range of models including simple lumped conceptual models to more complex physically based distributed models [*Beven*, 2011]. At continental to global scales, this relies on the GIMs, which are in turn, much more complex models that need a careful balance between accounting for the spatial heterogeneity of hydroclimatic processes and the computational burden associated with the multiplication of near-homogeneous areas. Also, differently from basin-scale hydrological models, which are routinely calibrated against observed river discharge, GIMs are usually not calibrated [*Müller Schmied et al*., 2014] and are instead tuned to set parameter values. For instance, for the Macroscale Probability-Distributed Moisture (MacPDM) GIM, tuning involves tests of precipitation data sets and potential evaporation calculations against long-term average runoff and long-term average within-year runoff patterns [*Gosling and Arnell*, 2011].

Following the climate community and programs like the Climate Model Intercomparison Project, e.g., phase five [*Taylor et al*., 2012] the hydrological community has started modeling experiments using different global impact models driven by the same climate forcing. The first such initiative was the Water Model Intercomparison Project (WaterMIP) project [*Haddeland et al*., 2011], since followed for example by the Inter-Sectoral Impact Model Intercomparison Project (ISI-MIP) project [*Warszawski et al*., 2014]. As a result, the scientific community has now easy access to many multi-impact model ensembles providing information on the possible projections in hydrological variables in the future for the world. Along with ease of access comes the danger of the data being used not appropriately, for example if some members of the ensemble are poor at reproducing some part of the hydrological processes, that could result in misleading interpretation of the projections if caution is not taken. This is because the global models used for experiments such as WaterMIP and ISIMIP have generally been developed for different purposes—e.g., water resource availability assessment (Global Water Availability Assessment (GWAVA), Water Global Assessment and Prognosis (WaterGAP)), carbon fluxes (Lund-Potsdam-Jena (LPJ)), and water and energy fluxes (Joint UK Land Environment Simulator (JULES))—using different protocols for their parameterization and error reduction, hence likely to have been tested differently for reproducing different processes. Moreover, each model run can use a different setup which is generally not fully published, and it is never guaranteed that the same setup used to produce the result published in a paper has been used for another simulation. It might therefore not be appropriate to rely on previous assessment to evaluate the skill of a new ensemble. Furthermore, due to the scale and complexity of such global models, their parameterization requires a long process, much more complex than that required for catchment models. In particular, comprehensive sensitivity testing of all parameters is a very ambitious task seldom undertaken by developers. While not all model codes are available to the research community to use, it would require a huge (unrealistic) effort for someone not familiar with those models to undertake a uniform parameterization testing for all global impact models together.

To improve our confidence in the GIMs, namely, in climate impact studies, a necessary first step is the evaluation of the models' ability to reproduce the observational records. On this issue, *Prudhomme et al*. [2011] emphasized how an appraisal of the performance of large-scale models in replicating historical hydrological extremes is a necessary precursor to assessing the suitability of such models for projecting characteristics of hydrological extremes into the 21st century.

Model intercomparison frameworks like the aforementioned Water Model Intercomparison Project (WaterMIP) provide the opportunity to compare model simulations from a number of GIMs all driven with the same meteorological forcing: the Watch Forcing Data (WFD) [*Weedon et al*., 2011]. The WaterMIP GIMs have been evaluated with respect to low, medium, and high flows in a number of studies

Table 1. Main Characteristics of the Models Used in This Study (After *Haddeland et al.* [2011])

	Model	Time Step	Meteorological Forcing Variables[a]	Energy Balance	Evapotranspiration Scheme[b]	Runoff Scheme[c]	Snow Scheme
GHMs	WaterGAP	Daily	P, T, LW_{net}, SW	No	Priestley-Taylor	Beta function	Degree-day
	LPJmL	Daily	P, T, LW_{net}, SW	No	Priestley-Taylor	Saturation excess	Degree-day
	MPI-HM	Daily	P, T	No	Thornthwaite	Saturation excess/beta function	Degree-day
	GWAVA	Daily	P, T, W, Q, LW_{net}, SW, SP	No	Penman-Monteith	Saturation excess/beta function	Degree-day
	MacPDM	Daily	P, T, W, Q, LW_{net}, SW	No	Penman-Monteith	Saturation excess/beta function	Degree-day
LSMs	HTESSEL	1 h	R, S, T, W, Q, LW, SW, SP	Yes	Penman-Monteith	Infiltration excess/Darcy	Energy balance
	JULES	1 h	R, S, T, W, Q, LW, SW, SP	Yes	Penman-Monteith	Infiltration excess/Darcy	Energy Balance
	MATSIRO	1 h	R, S, T, W, Q, LW, SW, SP	Yes	Bulk formula	Infiltration and saturation excess	Energy balance
	Orchidee	15 min	R, S, T, W, Q, LW, SW, SP	Yes	Bulk formula	Saturation excess	Energy balance

[a] R = rainfall rate, S = snowfall rate, P = precipitation (rain or snow distinguished in the model), T = air temperature, W = wind speed, Q = specific humidity, LW = longwave radiation flux (downward), LWnet = longwave radiation flux (net), SW = shortwave radiation flux (downward), SP = surface pressure.
[b] Bulk formula: Bulk transfer coefficients are used when calculating the turbulent heat fluxes.
[c] Beta function: Runoff is a nonlinear function of soil moisture.

[*Gudmundsson et al.*, 2012a; *Haddeland et al.*, 2011; *Prudhomme et al.*, 2011; *Stahl et al.*, 2012; *Tallaksen and Stahl*, 2014; *Van Loon et al.*, 2012], showing considerable variability in the magnitude and timing of the components of the hydrological cycle. Notably, all of these studies focused on Europe, despite the global coverage of the WaterMIP data set. Little is known about the skill of these models in reproducing the hydrological processes for other regions of the world. In this study, we address this gap in our knowledge by aiming to examine the capability of nine GIMs to reproduce key features of the hydrological regime, including low, medium, and high flows over the central United States (defined as the region between 36°N to 49.5°N and −105°E to −80°E): a region that experiences a wide array of hydrological extremes, with the 1993, 2008, 2013, and 2014 flooding events and the 1988 and 2012 droughts representing some of the most recent extremes, and where water availability is critical for agricultural production.

2. Data and Methods

In this study a first level of analysis uses a larger streamflow data set to verify whether the models are able to capture overall trend patterns of regional hydrology and two specific extreme events (1988 drought and 1993 flood), and a second level uses a smaller set of gauges (whose catchment have comparable size with the grid cells) to evaluate model performance matching observed and modeled data at the gauge-grid cell scale. This framework was chosen to ensure a first level of analysis with a sufficient number of streamflow gauges for spatial representativeness in the trend (section 2.4) and extreme event (section 2.5) comparison and a robust second level of analysis on carefully selected pairs (section 2.6).

The rationale behind this choice is that model evaluations must deal with a misalignment between modeled and observational data: as pointed out by other authors [e.g., *Gudmundsson et al.*, 2012a], large-scale hydrological models are not designed to model runoff at the catchment scale and interpreting localized model performance by comparing it with observed data may yield misleading results. Modeled data are systematically distributed in grid cells over the study region at a given spatial resolution, while the observational records do not have the same homogeneous coverage. Also, stream gauges provide an integrated measurement over a catchment [e.g., *Hannah et al.*, 2011], while the runoff information provided by the models represents values uniformly distributed over grid cells.

2.1. Simulated Data

We use daily total (surface plus subsurface) unrouted runoff outputs from nine GIMs created as part of the WaterMIP project. WaterMIP comprises both land surface models (LSMs) and global hydrological models (GHMs). As mentioned before, the key difference between these two types of models is whether they solve at the land surface both the water and the energy balances (LSMs) or only the water balance (GHMs). These models vary in structure and parameterization; we provide a brief overview of the set of models in Table 1 (for a comprehensive description of the characteristics see *Haddeland et al.* [2011]). All of the global models were run over the period 1963–2001 (except GWAVA: 1963–2000) at a spatial resolution of 0.5 decimal degrees and forced by the same meteorological input data: Watch Forcing Data (WFD). The WFD [*Weedon et al.*, 2011] was derived from the ERA-40 reanalysis [*Uppala et al*., 2005], interpolated to a 0.5° resolution and bias-corrected based on the Climate Research Unit data (of the University of East Anglia) and the Global Precipitation Climatology Centre version 4 data.

The models (see Table 1) vary substantially in the parameterizations of evaporation and runoff and do not all use the same input variables or model time steps (in particular, all GHMs are run at a daily time step whereas LSMs are run at a subhourly time step).

As noted in the Introduction, in contrast to basin-scale hydrological models, which are routinely calibrated against observed river discharge, GIMs are usually not calibrated [*Müller Schmied et al*., 2014]. With the exception of WaterGAP, none of the models used in this study were calibrated specifically for the WaterMIP experiment, although they may have been calibrated for previous studies [*Haddeland et al*., 2011]. The GIMs use their default soil and vegetation information derived from mapped land properties (e.g., soil texture and vegetation density) [*Gudmundsson et al*., 2012a], and no attempt was made to standardize these parameters [*Haddeland et al*., 2011]. WaterGAP underwent a limited calibration procedure using local measured streamflow data (for details see *Hunger and Döll* [2008]).

2.2. Observations

We use daily discharge data covering the 1963–2001 period from 252 stream gauging stations (Figure S1a and Table S1 in the supporting information) as reference data set. The size of these catchments varies, with drainage areas ranging from 64 to 1,350,000 km^2, with a majority (80%) with area up to 7000 km^2 (see Figure S1b, while the catchment boundaries are shown in Figure S1c). Because no land use changes or water management interventions are accounted for in the modeled data, we selected these 252 gauges from the Hydro-Climatic Data Network (HCDN). This data set was introduced in 1992 and updated in 2011 [*Whitfield et al*., 2012] as a subset of U.S. Geological Survey (USGS) streamflow gauging stations with historical streamflow data responsive to climatic variations, so relatively free of anthropogenic influences such as dam impoundment, regulation, and wide-scale urbanization (although minor impacts may still be present, e.g., land use change).

2.3. Hydrological Indices

We aim to analyze changes in discharge over different parts of the flow regime (including high, medium, and low flows). The central United States is a region marked by a high flow season mostly from April to July [e.g., *Villarini et al*., 2011] and a low flow season usually from September to February. We focus on different hydrological indices extracted from daily discharge time series over the period 1963–2001 (except for GWAVA, for which data were available for 1963–2000) for both observed (252 gauges) and modeled (1350 grid cells) data. The hydrological year is January–December for high and medium flow indices and April–March for low flow indices. We use three magnitude and three timing indices: (1) annual maximum flow (AMax: a record of the largest daily discharge value for every year), (2) annual medium flow (AMed: a record of the median daily discharge value for every year), and (3) annual minimum flow (AMin: a record of the smallest daily discharge value for every year). Three timing indices were used to gain a basic understanding of whether the models are able to capture the timing of flooding, medium, and drought discharge: (1) annual maximum date (AMaxDate: the day of the year in which the largest daily discharge value occurs for every year); (2) medium flow date (V50Date: the day of the year by which half of the annual total discharge volume has occurred); V50Date follows the concept of “center of mass” timing proposed by *Stewart et al*. [2005], and also used, for instance, in *Moore et al*. [2007]; and (3) drought start date (VDef10Date: the day of the year by which 10% of the annual volume deficit has occurred). The threshold used to define the VDef10Date corresponds to the 20th quantile of the time series; following the center of mass concept over the volume deficit (as, for instance, in *Giuntoli et al*. [2013], which provide a schematic of the index), the drought starts on the day the 10% of the annual volume deficit has occurred. The latter timing index poses some limitations in the presence zero (or very low values) rich time series for which the index cannot be extracted or there are too few threshold crossings over the time series to provide useful information. Therefore, if the index has insufficient nonzero values (at least 25 over 38) it is screened out (shown in grey on the maps). In this regard, it is worth noting that other studies have highlighted how low flow tractability can be problematic for GIMs. For instance, *Gudmundsson et al*. [2012a] found that the performance of this same set of GIMs decreased systematically from high (Q_{95}) to low (Q_5) runoff percentiles over Europe. The ensemble median of the

GIMs, calculated as the median of the single GIMs' index series, was added to complement the results and assess whether its results are more satisfactory than for any of the GIMs.

2.4. Trend Patterns in Hydrological Indices

A first step in our evaluation is geared toward the assessment of the skill of the GIMs in reproducing regional patterns of changes in the selected metrics, as well as their temporal evolutions. We examine temporal changes in discharge using the Mann-Kendall test (among others, consult *Helsel and Hirsch* [1992] for a description of this test). This is a nonparametric test (it does not require any distributional assumption) that allows the detection of monotonic patterns in the record of interest.

2.5. The 1988 Drought and the 1993 Flood

We selected two major hydrological extremes that occurred during the time of analysis (1963–2001), namely, the 1988 summer drought which affected most of the conterminous U.S. and the 1993 summer flood which affected the U.S. Midwest particularly. Both events have developed over a time span of approximately three months, from June to August, as reported by the NOAA Billion-Dollar Weather and Climate Disasters (https://www.ncdc.noaa.gov/billions/events). We thus assessed how well the GIMs captured these events by considering the mean summer runoff volumes (from 1 June to 31 August) of the year in which the event occurred and compared them to the mean summer runoff volumes over the whole time series. These differences are quantified using the following coefficient of variation (e.g., for the 1998 drought):

$$\mathrm{CV} = Q_{\mathrm{JJA}[88]} - Q_{\mathrm{JJA}[63-01]}/\sigma\left(Q_{\mathrm{JJA}[63-01]}\right)$$

We thus map this quantity to show whether the models indicate negative (positive) balances for drought (flood). In addition, we express the exceedance probability (p) by ranking the years based on their summer runoff volumes and compute the plotting position of the particular year event (1988 or 1993) with reference to the whole time series:

$$p = m/n + 1$$

where m is the rank position and n is the number of years in record.

2.6. Modeled-Observed Pairwise Comparison

We carry out a pairwise comparison between observed and modeled discharge using a subset of 128 nonnested gauges, which were selected within the 400 to 3500 km^2 catchment area range (Figure S1)—while the size of the model grid cells ranges depending on the latitude from approximately 2500 km^2 at 36°N to 2000 km^2 at 49.5°N. The selection of the pairs was carried out on a GIS using the streamgauges' catchment boundaries obtained from the National Weather Service (http://www.nws.noaa.gov/geodata/catalog/hydro/html/basins.htm): the grid cell corresponding to a given catchment was selected on the basis of centroid proximity. Priority was given to larger catchments (i.e., with area closer to the grid cells), and in case of more catchments overlapping over the same grid cell, the one that shared the majority of the area was selected. Because of the different units used for modeled and observed data—except for the timing indices (expressed in number of days from beginning of hydrological year)—the index series for the observed streamflow data were converted from cubic feet per second to millimeter of runoff per unit area per second.

The comparison is carried out first on the timing indices assessing the monthly frequency of occurrence; this is followed by analyses on all of the index series using three performance metrics: Pearson's correlation coefficient, computed to assess the similarity of the index series across pairs, with optimal value $R = 1$; the relative difference in standard deviation, computed to compare the amplitude of observed and modeled indices data, with optimal value $\Delta\sigma = 0$, root-mean-square error, computed to express the magnitude of the difference between observed and modeled indices series, with optimal value root-mean-square error (RMSE) = 0.

3. Results

3.1. Trend Patterns

Results related to the temporal change in AMax (Figure 1), AMed (Figure 2), and AMin (Figure 3) are presented through maps showing the sign and significance of the results of the Mann-Kendall test. Note that grid cells were grayed out when the total runoff was negative. These negative values can be achieved if, for instance, there is high evaporation and no sufficient precipitation to generate runoff, as seen for the WaterGAP and

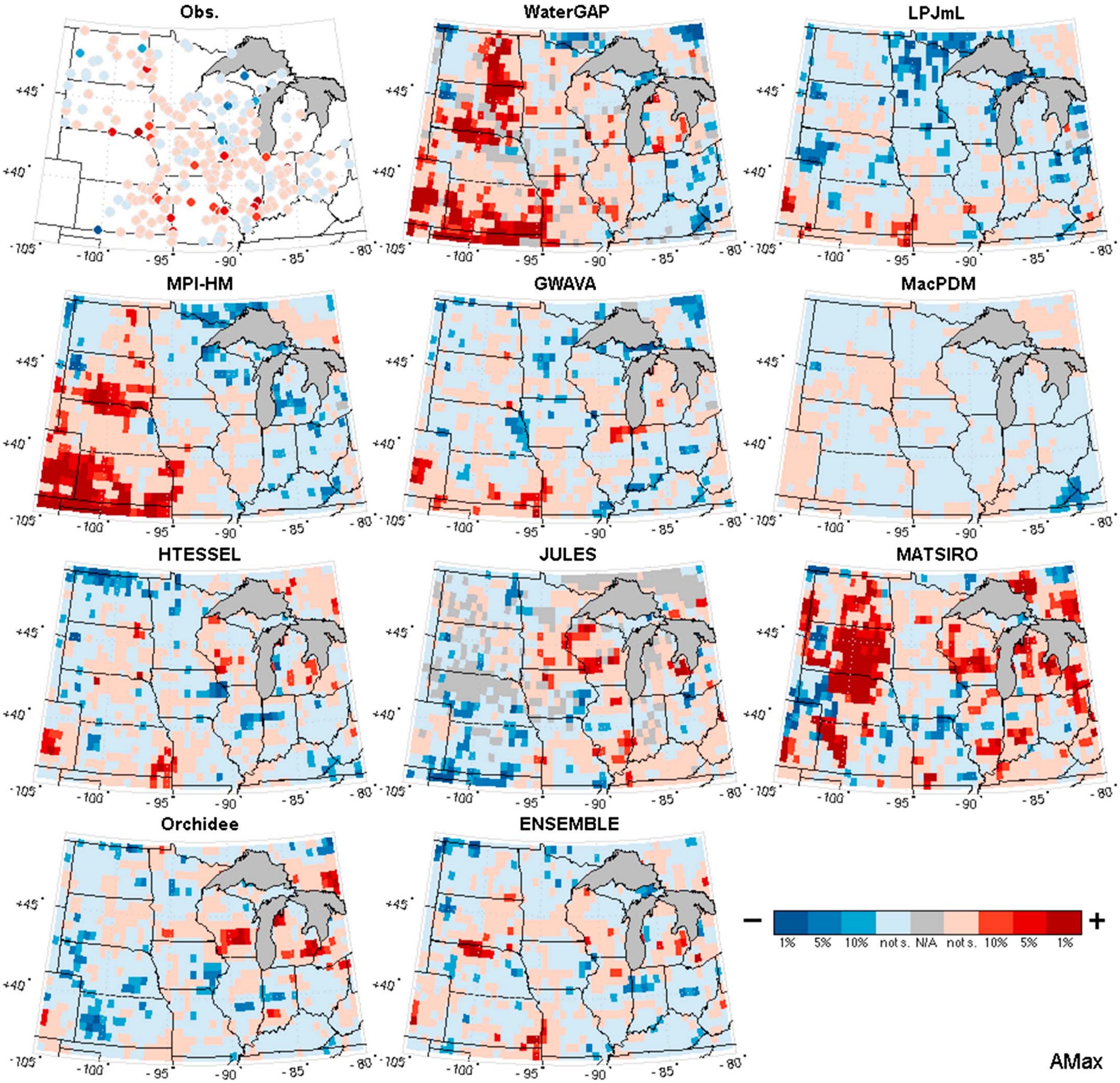

Figure 1. Trends in the annual maximum flow for (top left) observed data, the nine GIMs, and their ensemble median. Negative trends are shown in blue and positive trends in red, with three levels of significance (1, 5, and 10%) from pale (not significant) to dark (significant at the 1% level).

JULES [*Döll and Schmied*, 2012; *Williams and Clark*, 2014]. Grid cells were also grayed out when runoff was unavailable—for the Great Lakes (WaterGAP, LPJmL, Max Planck Institute Hydrology Model (MPI-HM), and GWAVA) or when the hydrological index tested had null variance (e.g., all annual minima equal to zero). Also, note that there are very few streamflow gauges in the southwestern part of the study region. While there are a number of USGS stream gaging stations, a very small number are included in the HDCN, mostly because of large water withdrawal for agriculture [e.g., *Rasmussen and Perry*, 2001].

The annual maximum index based on the observations (Figure 1, top left) shows a weak tendency toward increasing trends over most of the region, although the trends are generally not significant at the 0.1

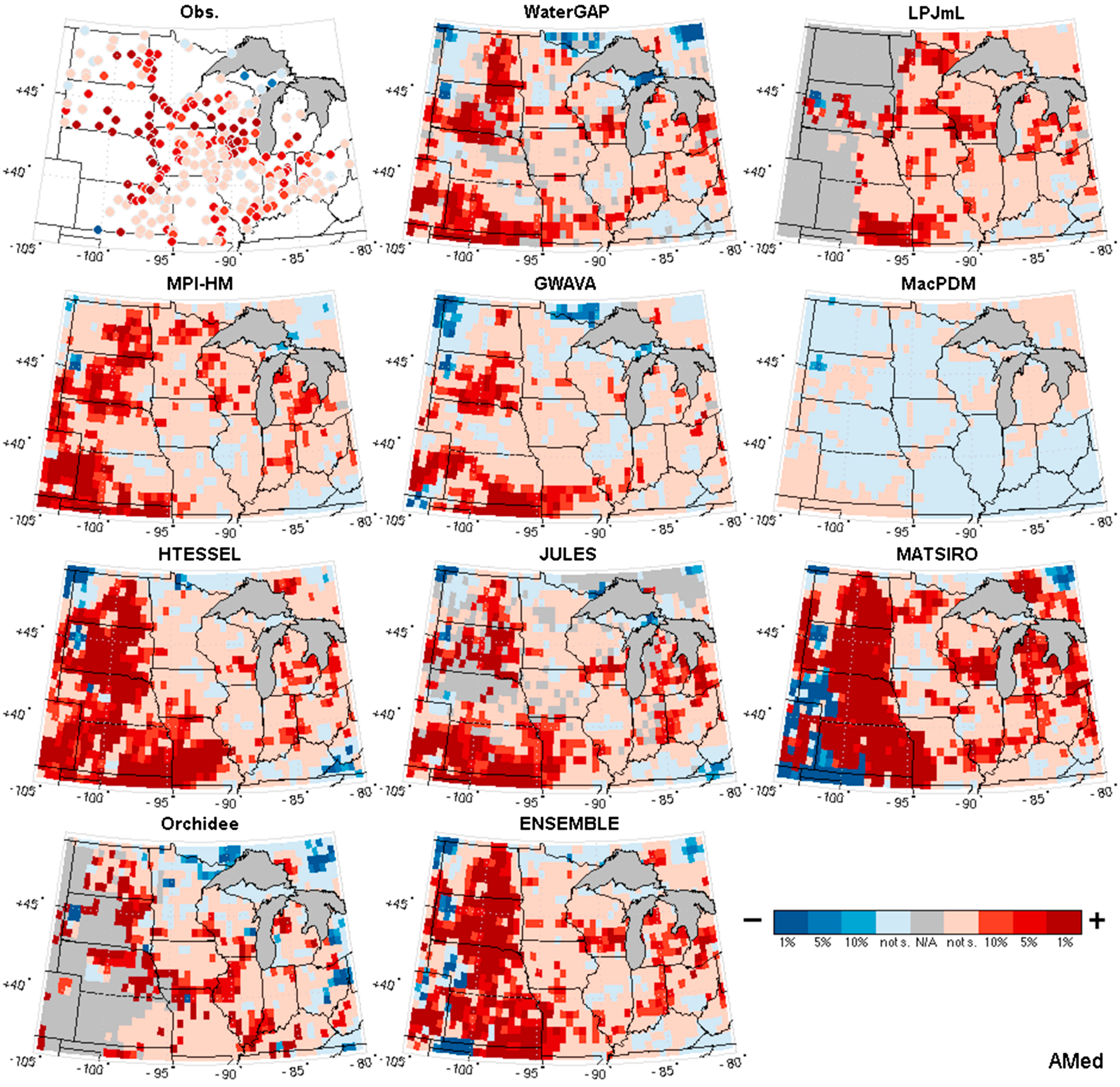

Figure 2. Same as Figure 1 but for annual medium flow.

significance level. These results are consistent with what discussed in the literature [e.g., *Hirsch and Ryberg*, 2012; *Peterson et al.*, 2013; *Villarini et al.*, 2011; *Vogel et al.*, 2011; *Mallakpour and Villarini*, 2015] where there is not a very strong indication of changes in extreme discharge over this area, but more of a tendency toward increasing trends. For the GIMs, MacPDM depicts a rather muted signal with virtually no significant trends over the entire region. In comparison, the remaining models show stronger patterns of change. In particular, WaterGAP, MPI-HM, and Minimal Advanced Treatment of Surface Interaction and Run-Off (MATSIRO) yield spatial patterns that more closely resemble the observations, with an even stronger signal of change than observed. Most of the models indicate a decreasing trend in northern Minnesota that could not be compared with the observations due to the lack of stream gaging stations in the area. The lack of observational records

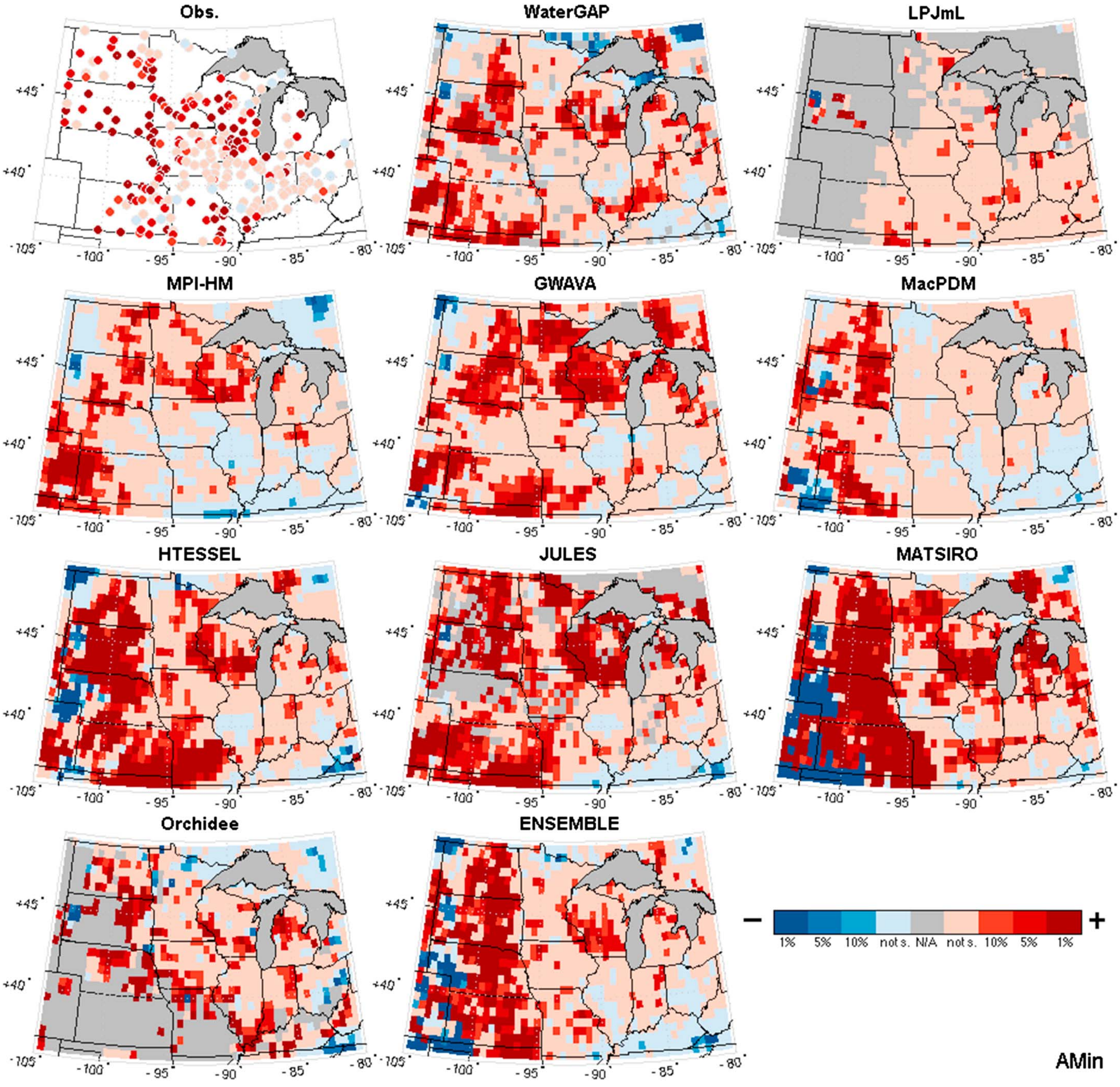

Figure 3. Same as Figure 1 but for minimum flow.

holds true for the area including Nebraska and Kansas, for which the models suggest increasing trends in annual maximum daily discharge. The models GWAVA, Tiled ECMWF Scheme for Surface Exchange over Land (HTESSEL), JULES, and Orchidee show a generally noisier signal with both positive and negative trends over the region of study.

Trends in medium (Figure 2) and minimum (Figure 3) discharge show a much clearer pattern than for the annual maximum daily series. These results are consistent with published work [e.g., *Douglas et al.*, 2000; *Lins and Slack*, 1999, 2005; *McCabe and Wolock*, 2002], in which most of the statistically significant increasing trends were detected for low to moderate quantiles, and much fewer when dealing with annual maximum discharge.

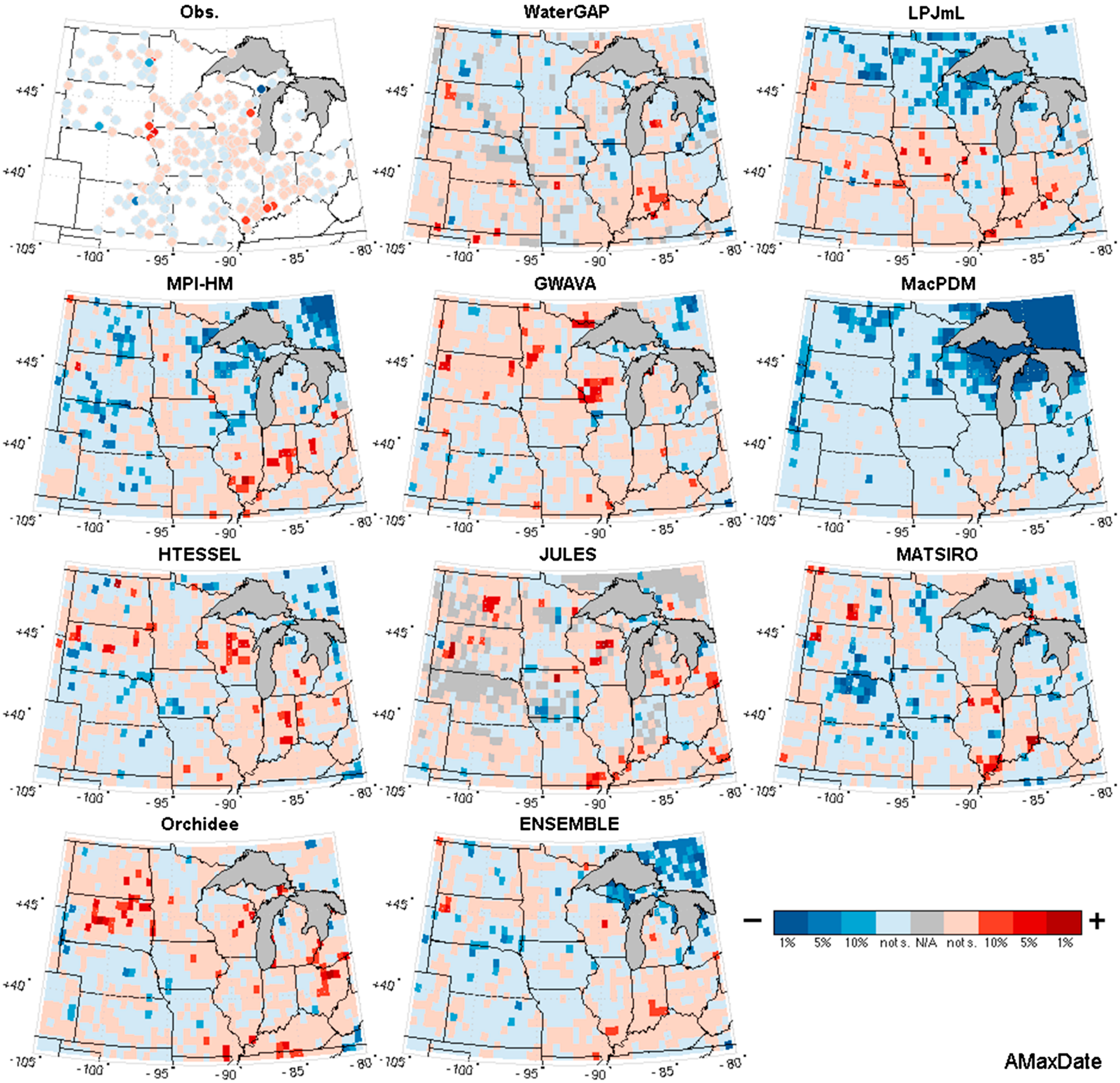

Figure 4. Same as Figure 1 but for annual maximum flow date (positive trends indicate events occurring later, negative trends earlier).

Trends in observed annual minimum indicate strong and highly significant (*P* values generally <0.01) increasing trends over most of the region, with the exception of the southeastern part of the domain (weaker signal). Overall, the models capture well this increasing pattern. In particular, the LSMs (HTESSEL, JULES, and MATSIRO) show strong increasing trends that are also detected, although not as strongly, in the GHMs (WaterGAP, GWAVA, and MPI-HM) and to a lesser extent in MacPDM (positive significant detections are limited to the western part of the domain). The models LPJmL and Orchidee have a substantial number of grid cells screened out (gray), where the annual minimum is equal to zero over the 38 years considered. This behavior results in a large part of the pixels being removed from the analysis in south-west for Orchidee, and in the west and the north-east for LPJmL. In the unmasked areas, Orchidee reproduces well

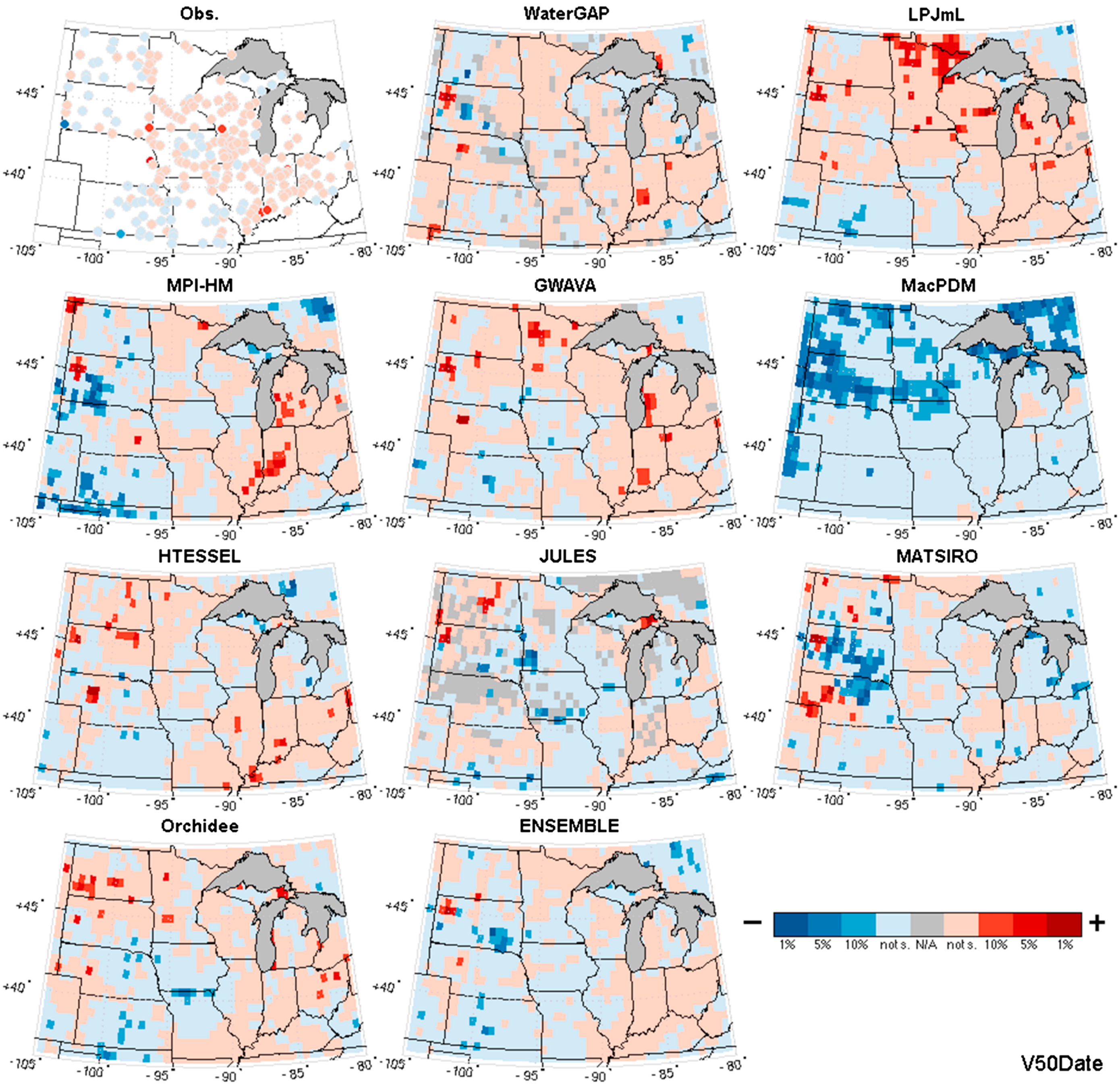

Figure 5. Same as Figure 1 but for annual medium flow date (positive trends indicate events occurring later, negative trends earlier).

the spatial signal patterns with positive trends, whereas LPJmL shows no significant detections (this is also the case for MacPDM over the same area). Thus, LPJmL and MacPDM do not seem to capture the overall trend in runoff annual minima as well as the other GIMs.

Trends in medium flow (Figure 2) are broadly similar to those for the annual minimum flow, with most GIMs capturing the observed overall increasing signal. In contrast with the other GIMs, MacPDM has virtually no significant trends. Although less than for the AMin, LPJmL, and Orchidee have grid cells screened out even for the medium flow. This is rather surprising because it indicates that at least half of the days every year have daily discharge equal to zero. At this stage, it is unclear what the issues with these two models are, although this issue was also noted by *Gudmundsson et al*. [2012a] where the two GIMs have constant low values of

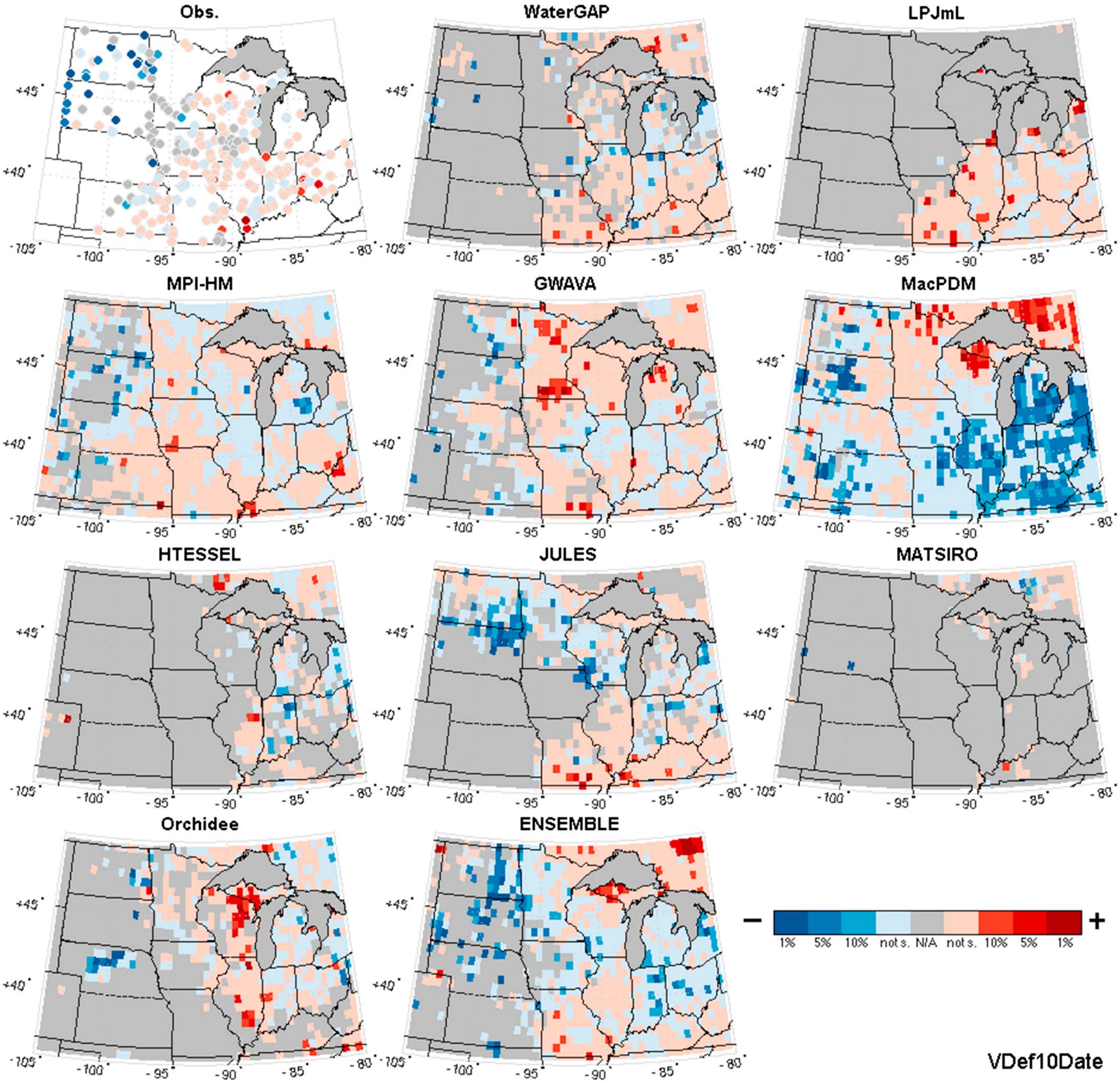

Figure 6. Same as Figure 1 but for annual volume deficit 10% date (positive trends indicate events occurring later, negative trends earlier).

interannual variability at low percentiles (i.e., Q_5 and Q_{25}) and by *Prudhomme et al*. [2014] where LPJmL displays a similar behavior in the runs of the ISI-MIP experiment.

We focused also on the timing of high (AMaxDate; Figure 4), medium (V50Date; Figure 5), and low flows (Vdef10Date; Figure 6) to aid inference of the discharge-generating processes over this region. The observations do not point to a change in the seasonality of high flow or medium discharge, with no statistically significant (at the 0.1 level) trends. The lack of a clear spatial pattern and significant trends in the date of annual maxima is reproduced by most GIMs (WaterGAP, GWAVA, HTESSEL, JULES, MATSIRO, and Orchidee; Figure 4). However, decreasing trends are simulated in the north/north-east part of the region by three of the GHMs (LPJmL, MPI-HM, and MacPDM) and by the ensemble median. These results would indicate an

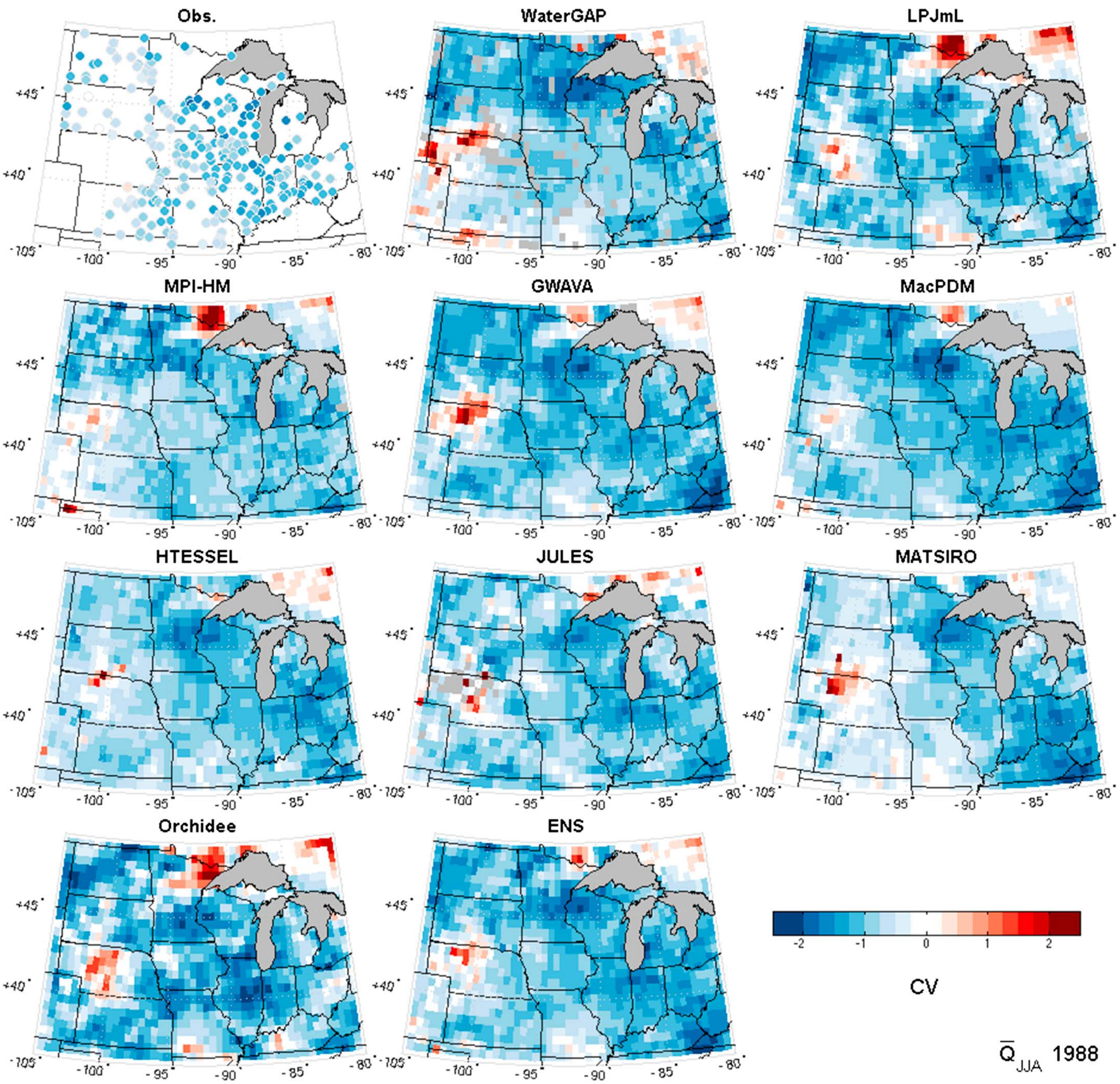

Figure 7. The 1988 drought coefficient of variation for (top left) observed data, the nine GIMs, and their ensemble median. Negative CVs are shown in blue and positive CVs in red (negative CVs indicate 1988 summer mean runoff smaller than mean 1963–2001 summer mean runoff).

earlier occurrence of annual peaks, potentially linked to an earlier melting of the snowpack. While this finding would be consistent with increasing temperatures [e.g., *Villarini et al*., 2013], it is not picked up in observational records at the 0.1 significance level. The medium flow date (Figure 5) shows very few trend detections for both the observed and the GIMs (including the ensemble median). A few models show areas with decreasing trends—as seen for the maximum flow date—especially in the north (MacPDM) and to a lesser degree in the west (MATSIRO and MPI-HM), while LPJmL shows an increasing trend in the north. Except for the marked decreasing pattern of MacPDM, the few hotspots seen in the other models are small and point to scarce detections and no clear overall pattern.

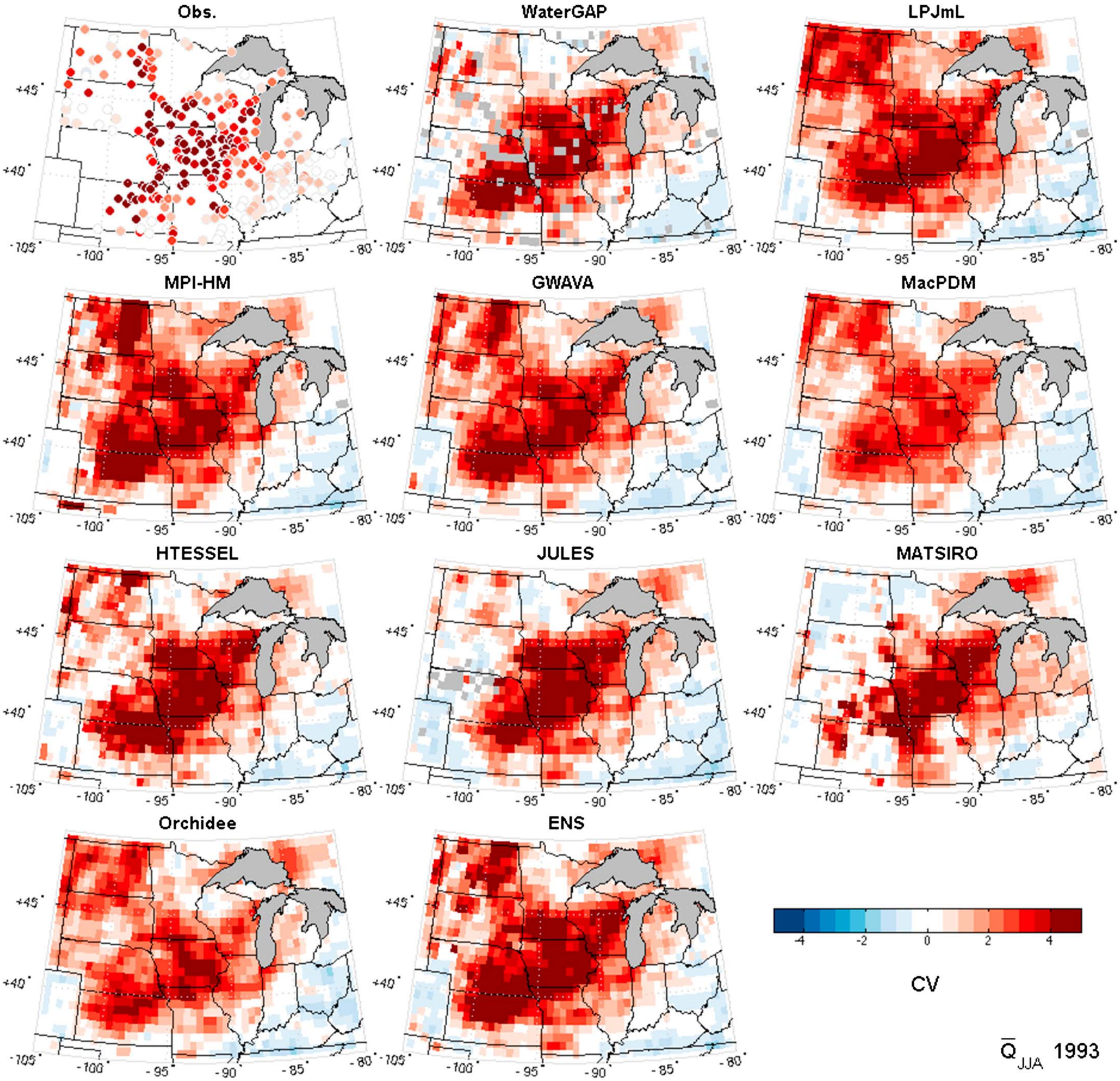

Figure 8. Same as Figure 7 but for 1993 flood (positive CVs indicate 1993 summer mean runoff larger than mean 1963–2001 summer mean runoff).

The drought start for observed data (expressed as volume deficit date; Vdef10Date) shows a few decreasing trends in the northwest (mostly North and South Dakota) and very few increasing trends in the southeastern part of the domain. This would hint at an earlier onset of the drought start in the northwest. The masking applied to the GIMs depends on whether the grid cells had sufficient nonzero values in the index (<25). In spite of the considerable masking, most models seem to match the weak pattern in the trends detected on the observed data, although MacPDM shows marked decreases in the southeast and an increase in the northeast. Finally, JULES and the ensemble median seem to capture well the light decreasing pattern present in the observations in the northwest part of the study region.

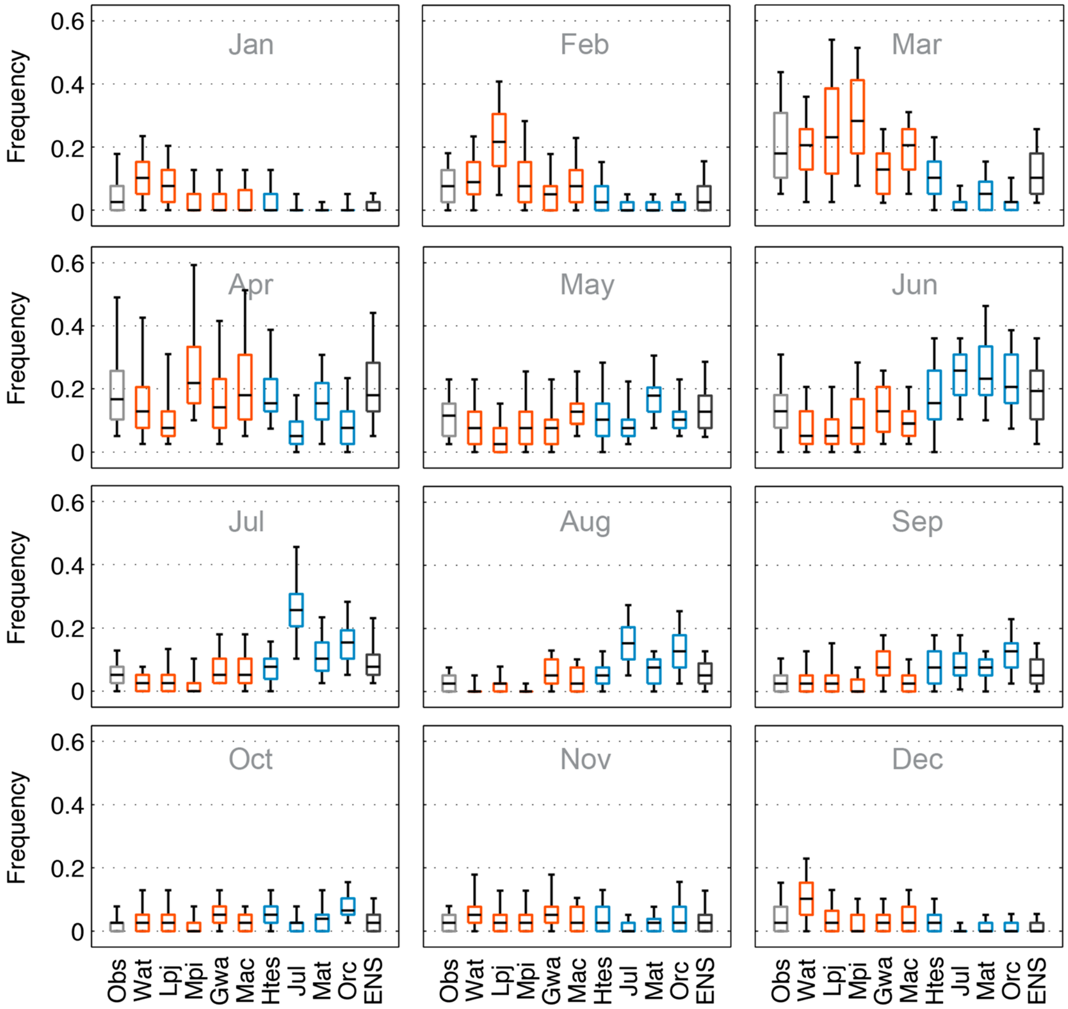

Figure 9. Frequency of occurrence of annual maximum flow per month for 128 gauges and corresponding grid cells (bar: median, box: interquartile range, whiskers: 10th and 90th percentiles). In light gray the observed records, in orange the GHMs, in blue the LSMs, and in dark gray the ensemble median.

3.2. The 1988 Drought and the 1993 Flood

All GIMs and the ensemble show good agreement with the observed data in capturing both the 1988 drought (Figure 7) and the 1993 flood (Figure 8). While the pattern is more evenly distributed for the 1988 drought, the 1993 flood appears intensified with a patch spanning from the southwest (Kansas) to the north east (Wisconsin) of the domain. The intensity of the variations is different for the two events; coefficient of variations (CVs) vary mostly between 0 and −2 for the 1988 drought and between 0 and 5 for the 1993 flood. This indicates that the 1993 summer flood volumes have a more pronounced departure from the whole period's summer volumes than the 1988 summer drought does. This is to be expected and can be explained by the more erratic nature of the flood runoff volumes compared to slower onset and development of the drought ones (whose values, differently from the flood, are bound to zero). The good performance in capturing these two events is confirmed by the exceedance probability maps (Figures S33 and S34), where, as expected, low probabilities result for the 1993 mean summer runoff and vice versa for the 1988. While all GIMs tend to capture the mean runoff differences with similar intensity and spatial pattern, MacPDM appears to capture the spatial pattern equally well, but with a weaker intensity with regards to the 1993 flood.

3.3. Modeled-Observed Pairwise Comparison

After considering the whole domain for the examination of trends of magnitude and timing indices, and the consideration of two particularly extreme events, we focus on a subset of stations to examine whether the

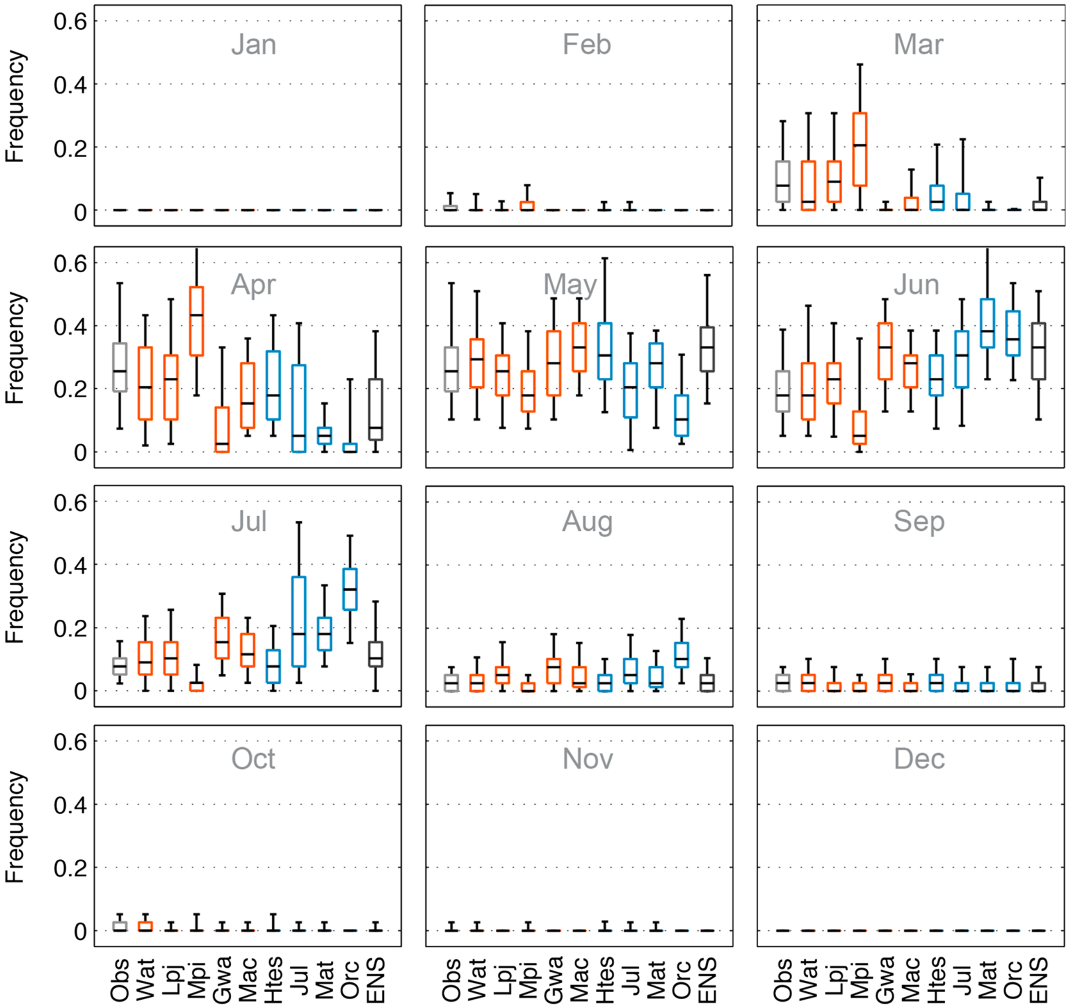

Figure 10. Same as Figure 9 but for annual medium flow.

models are able to capture the seasonality in these quantities. More specifically, we focus on 128 grid cells selected to correspond to 128 streamflow gauges.

3.3.1. Timing of Annual High, Medium, and Low Flows

For the entire region, the monthly frequency of occurrence of annual maxima and annual medium flows is shown as boxplots in Figures 9–11, while Figure 12 quantifies the differences in the median and the interquartile range with respect to the observations. It is worth clarifying that the boxplots summarize the results grouping outcomes from different regions and on a limited number of grid cells (128 of 1350). Maps of the monthly variability in the occurrence of annual maximum, medium, and low flows for each model are given in Figures S2–S31.

The observed annual maxima (Figure 9) occur mostly from March to June, with the highest frequency in April. This pattern is reproduced by the GIMs, but specific behaviors emerge depending on the nature of the model (LSMs versus GHMs). The GHMs tend to show a seasonality characterized by an enhanced frequency of occurrence of annual maxima about 1–2 months earlier than the observations, with medians that are closer overall to observed data. The LSMs, on the other hand, tend to exhibit a delayed seasonality (1–2 months later) and to show an overall greater discrepancy from observations. This pattern is shown very clearly in Figure 12 (top row), where the GHMs (WaterGAP, LPJmL, and MPI-HM) tend to overestimate count rates in AMaxDate occurrences from December to March and to underestimate them from April to September. Opposite to this pattern, the LSMs (JULES, MATSIRO, and Orchidee) tend to underestimate count rates from February to April and to overestimate them from June to September. The spread (quantified interquartile range (IQR)) of the

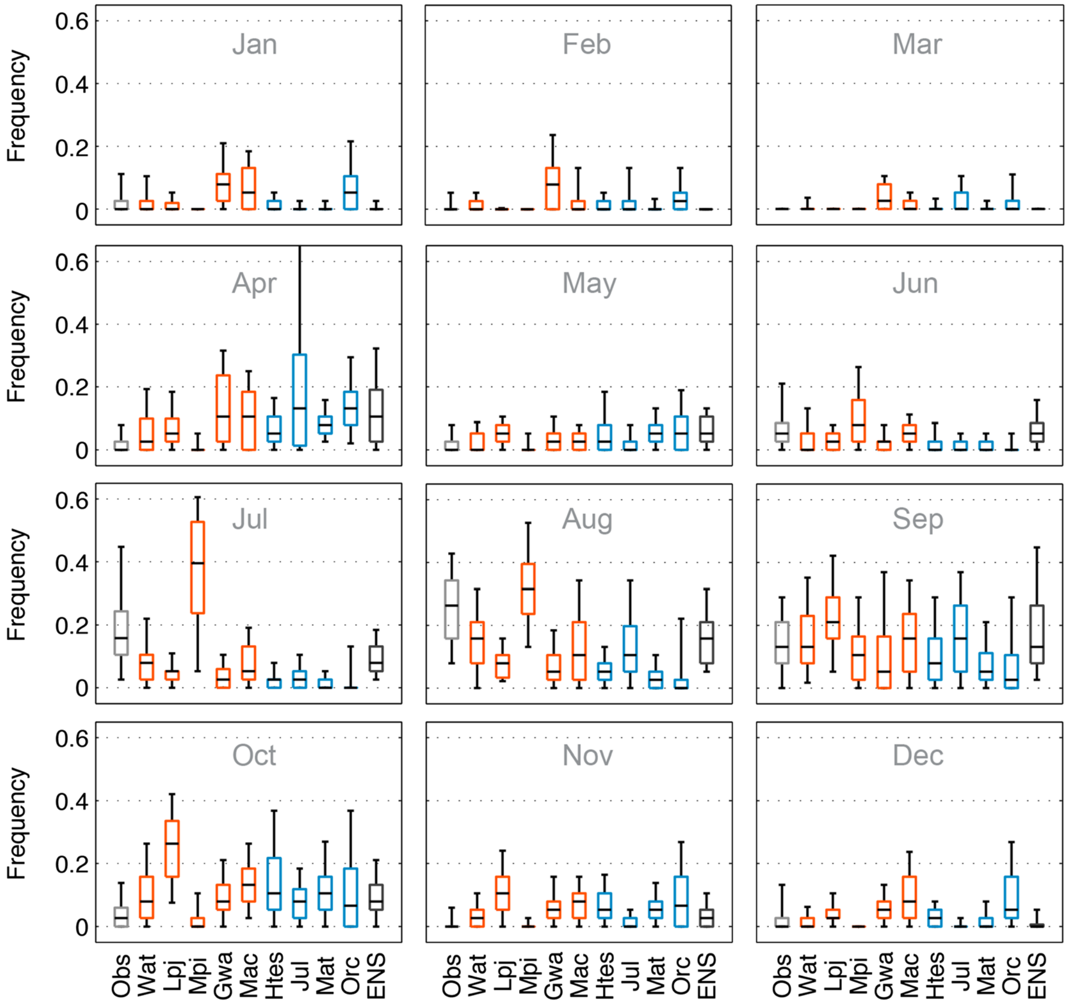

Figure 11. Same as Figure 9 but for annual drought start.

LSMs is higher when there is an overestimation of the count rates and lower in the case of underestimation, whereas the spread of the GHMs is generally closer to the observational one throughout the year. Between these two marked behaviors lay GWAVA, MacPDM (GHMs), and HTESSEL (LSM), which show the smallest differences from the observations both in the median and IQR.

The observed data indicate that the V50Dates occur from March to June with the highest counts in June (Figure 10). Few or no events are counted from September to February, and this is captured unanimously by all the models. For March to August, GHMs tend to capture better the timing of the medium flows than the LSMs, although there are some discrepancies among these models. More specifically, WaterGAP, GWAVA, and MacPDM underestimate the count rates in V50Date occurrences during March and April, while GWAVA and MacPDM also overestimate from May to August; LPJmL and MPI-HM underestimate them in late spring. With the exception of HTESSEL, which captures rather well the timing throughout the entire year, for the LSMs there is a marked underestimation during the spring (March to May) and an overestimation in the summer (June to August). The LSMs are strikingly not in line with the observations, and they appear to be out of phase with a lag of 1–2 months. Figure 12 (middle row) shows this phase shift for which the largest differences in the median and the IQR appear for the LSMs JULES, MATSIRO, and Orchidee and to a smaller extent for the GHMs MPI-HM and GWAVA.

The drought starts (Vdef10date) in observed data show few occurrences in the spring (April–May) and an increasing frequency in the summer, peaking in August and decreasing in early fall (September–October), with virtually no occurrences in the winter from November to March (Figure 11). The GIMs ability to

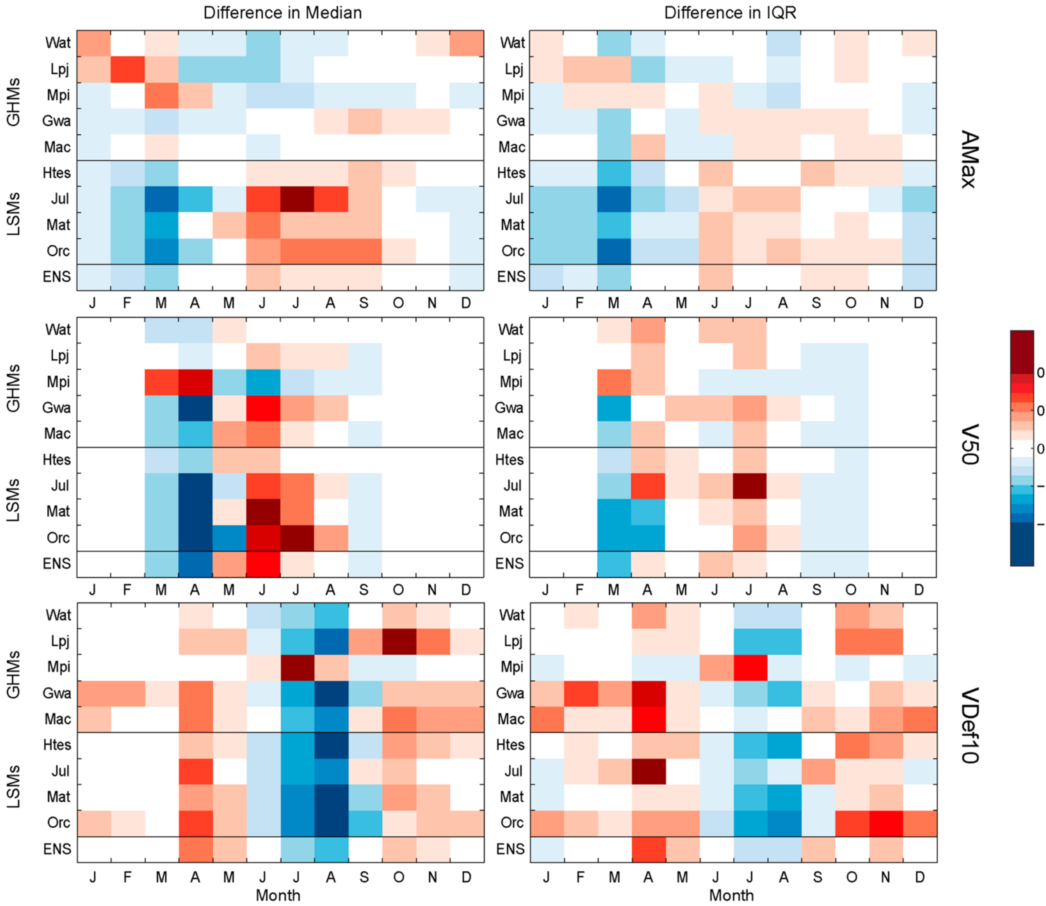

Figure 12. Occurrence of (top row) annual maximum, (middle row) annual medium flow, and (bottom row) annual drought start events per month (as seen in Figures 9–11): difference in (left column) the median and (right column) the interquartile range IQR of the models from the observations—red, overestimation; blue, underestimation.

reproduce ground observations is weak, highlighting the difficulty in capturing the timing of low flows with respect to high and medium flows. For instance, for the two previous indices (Figures 9 and 10), months with no occurrences were broadly well reproduced by the majority of the GIMs, while for the drought start some GIMs show considerable frequencies, especially in the winter as opposed to the frequencies of the observations that are near zero; there is also a less pronounced homogeneous response per type of GIM seen thus far. With the exception of MPI-HM, which seems to follow the most closely the observed results, all of the other GIMs show noticeable fewer counts in the summer when counts are high. The situation changes in September, when GIM counts increase and tend to decrease in the fall at a much slower rate than the observed data. This lag seems to indicate that GIMs tend to capture the drought onset later in the year, with approximately a 1–2 month delay. In addition, there are higher frequencies in winter and spring. This is visible in Figure 12 (bottom row), where there is clear marked underestimation of the drought start in the summer (especially July and August) and an overestimation in spring and fall.

3.3.2. An Assessment of the GIMs' Performance

Figure 13 summarizes the results of the performance achieved by the GIMs in the pairwise comparison for the hydrological indices from the streamflow gauges and from the corresponding grid cell. The first index, Amax, depicts a performance that is fairly homogeneous across the GIMs. The main differences are for the *R* coefficient, according to which GHMs perform slightly better than the LSMs. For the annual medium discharge performances improve in all metrics compared to the Amax: the GIMs' correlation to the observed data improves noticeably, with *R* values approaching 1; the $\Delta\sigma$ are closer to zero and their spreads decrease; the RMSE values show that GIMs are closer to the observations. The other end of the hydrological regime, the annual minima, seems to perform better in the RMSE and *R* correlation than the annual maximum, but results within models in the $\Delta\sigma$ can differ considerably in the spread.

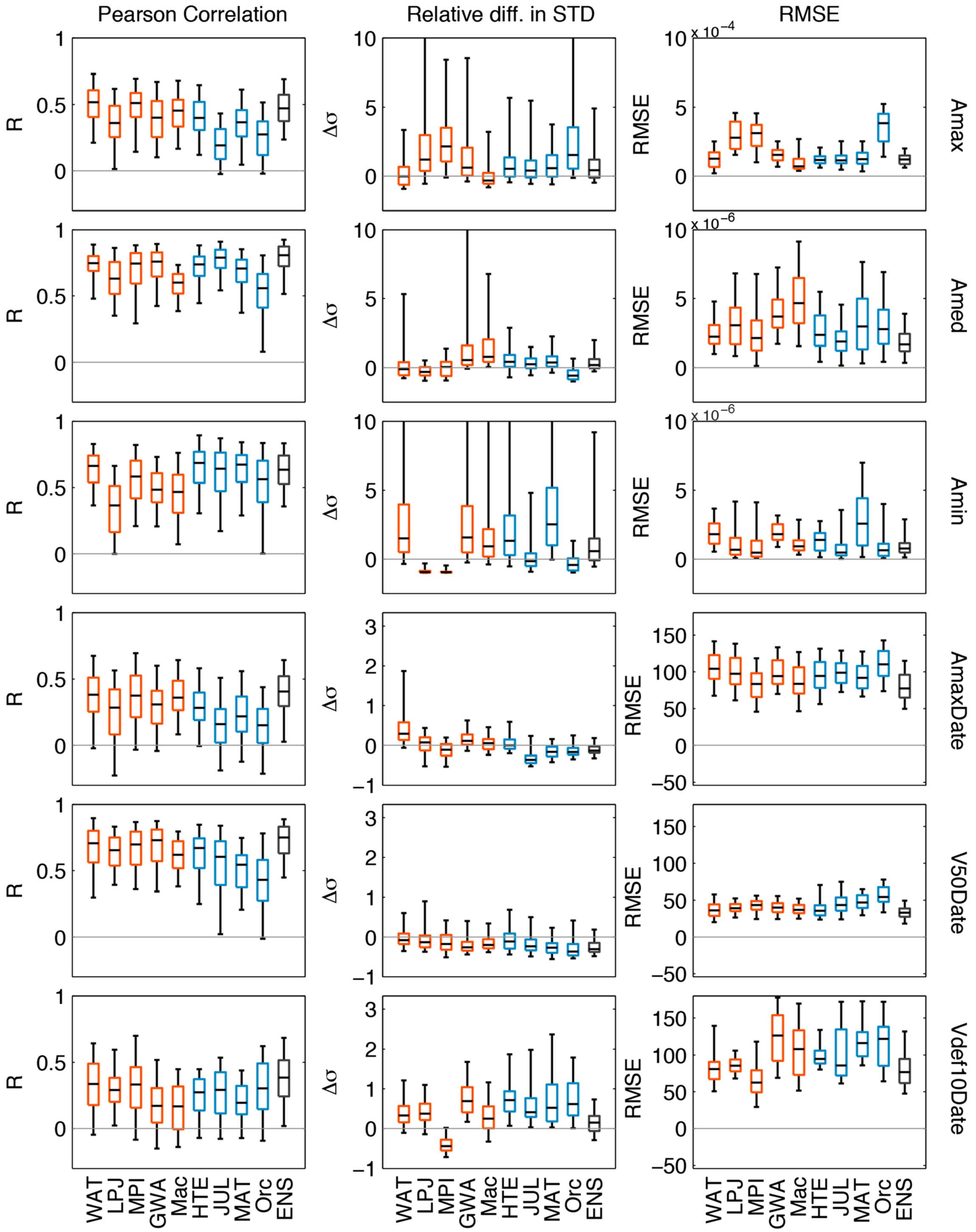

Figure 13. Performance metrics (in column: Pearson's R correlation coefficient, relative difference in standard deviation $\Delta\sigma$, RMSE) on the pairwise comparison observed-modeled (128 points) for the six hydrological indices (in row). For the boxplots: bar, median; box, interquartile range; whiskers, 10th and 90th percentiles. Notice that the vertical scales are different for (middle column) $\Delta\sigma$ and (right column) RMSE.

The results for the annual medium discharge have less pronounced variability in $\Delta\sigma$. This can be due to the description of the central part of the hydrological regime, as opposed to intrinsecally more erratic nature at the tails (Amax and Amin). Similarly, lower values of correlation (R) of Amax compared to Amin may be partly owed to Amax's more erratic behavior, while Amin is bounded below at zero. It should be noted that the index series comparison modeled-to-observed is based on approximately 39 points (1963–2001) for high and medium flows and on 38 (1963–2000 as the hydrological year starts in April) for low flows. Also, while computing the metrics, a year with missing value found in one of the two series is excluded from both series.

The following three rows in Figure 13 describe the timing of high, medium, and low flows indices expressed in number of days from the beginning of the hydrological year. In general, similar to the previous three indices, the ensemble median seems to outperform individual GIMs, and the medium flow (V50Date) is the index that is the closest to the observations. Focusing on the correlation coefficient, the second best index is the annual maximum flow (AmaxDate) with the GHMs performing better than the LSMs, followed by the annual drought start (Vdef10Date). Similar to V50Date, the $\Delta\sigma$ nears zero for most of the GIMs for AmaxDate. The same is not true for the Vdef10Date, which have higher values and larger spreads. The RMSE stays below 50 for V50Date, and around 100 for AmaxDate and Vdef10Date, although the latter shows stronger variations from GIM to GIM, including in the spread. It should be noted that results for Vdef10Date tend to include fewer than 128 pairs because the presence of zeros in the index series (the threshold was not always crossed) affecting the pairwise comparison: series with less than 25 values different from zeros were excluded. The GIMs using fewer pairs are LPJmL and MATSIRO (47 pairs), followed by HTESSEL (71), Orchidee (78), with the remainder of the GIMs having between 109 and 124 pairs.

4. Discussion and Conclusion

The aim of this paper was to assess how well the regional hydrology of the central United States (based on observations at 252 reference gauges from 1963 to 2001) was reproduced by a set of nine global impact models from the WaterMIP Project and their ensemble medians. The focus was on the examination of a number of discharge indices related to high, medium, and low flows, as well as the seasonality and timing of the flow regime.

In our model-observation comparison, there are few elements that we need to keep in mind when interpreting these results. The spatial resolutions of the models and observed records used as reference are not the same. The models do share a historical forcing (the Watch Forcing Data (WFD)) that has been provided globally and whose quality can vary depending on the region. However, our study region lacks high-elevation features, which typically have a negative effect on the quality of the forcing, and, more importantly, the scale at which we operate for the trend detection is sufficiently large to allow for a comprehensive comparison of the patterns, while for the pairwise comparison analysis, the observed data set is reduced using only catchment of comparable size with the grid cell.

To date, the WaterMIP GIMs have been used in other studies [e.g., *Gudmundsson et al*., 2012a; *Prudhomme et al*., 2011; *Stahl et al*., 2012] comparing their control period with observed data over parts of Europe. A general conclusion was that the models tend to capture the interannual variability of high, medium, and low flows well. All of these studies show that simulated runoff can vary substantially depending on the GIM, as every model has different characteristics in the way it simulates the different components of the water cycle. The type of flow (high or low) also plays a role: *Gudmundsson et al*. [2012a] show that for low runoff percentiles the performance of the models decreases, reflecting the uncertainty associated with the representation of the hydrological processes (e.g., the depletion of soil moisture storage). The same authors confirm the results by *Haddeland et al*. [2011] on MATSIRO's propensity to predict less seasonal variation in runoff than the other models. This is owed to a deep groundwater reservoir that buffers the timing of runoff, in turn leading to an underestimation of the magnitudes and to delays of the high flows peaks. Moreover, *Prudhomme et al*. [2011] focused on three WaterMIP GIMs and showed that WaterGAP is the model that best reproduces the regional characteristics of high and low flow events in Europe, while JULES and MPI-HM tend to have a slow and fast responding runoff, respectively. *Tallaksen and Stahl* [2014] focused on droughts (using seven WaterMIP GIMs) and also suggested that WaterGAP and GWAVA are better at capturing hydrological droughts over Europe.

The findings outlined above are generally consistent with our study: most of these GIMs are able to reproduce the spatial trends in the observational records over the central United States. However, a new element in our results is the clear dichotomy between LSMs and GHMs, which is reflected in the ability of each model to capture the timing of maximum and medium flows. The LSMs are less capable of capturing the timing exhibited in the observed data than the GHMs. For the annual maximum flow, GHMs tend to overestimate frequencies in the winter and to underestimate them during spring and summer, while the opposite is true and more marked for the LSMs. For medium flow, a strong underestimation of the frequencies is shown for the LSMs in the spring and an overestimation in the summer, while the GHMs are closer to the observations and show a less marked behavior in general. Though less marked, indications of similar behavior can be found in the works by *Haddeland et al*. [2011] and *Gudmundsson et al*. [2012b]. Over basins with a climate comparable to our study region (i.e., Northern Europe), *Haddeland et al*. [2011] showed that peaks occur earlier for GHMs than LSMs and linked this behavior to the snow scheme employed: the energy balance approach used by LSMs predicts reduced snow water equivalent values, leading to lower winter and spring runoff volumes than predicted by the degree-day approach used by GHMs. The snowy winters in the northern part of the central United States may explain the clear shift in the timing of high and medium flows yielded by GHMs and LSMs. It should be noted that as shown in Table 1, energy balance models (LSMs) comprise more forcing variables than degree-day models (GHMs) and are thus prone to additional associated errors.

The timing of low flows depicts a less marked behavior in terms of the type of GIM as seen for medium and high flows and also a poorer ability in capturing the frequencies of occurrence. In particular, GIMs' counts of drought start occur sporadically during seasons for which the observations display no counts. More importantly, during the summer, when observed data frequencies are high, GIMs tend to a generalized underestimation of the occurrences, and to an overestimation in the fall, when the results based on observations tend to decrease while the GIMs continue to have fairly higher rates. It is worth noting that the identification of drought start can be cumbersome when dealing with zero/very small values rich time series, by which some GIMs (e.g., LPJmL, MATSIRO, and Orchidee) are particularly affected (and to a lesser extent some streamflow gauges in part of the study domain). The problem is present even when choosing large thresholds quantiles, because for those grid cells/gauges whose runoff tends to plateau over most of the year and have an isolated very large peak, the threshold crossing may not occur every year (i.e., metric is not computed). This underlines the aforementioned increased difficulty of the GIMs to describe the lower tail of the runoff spectrum and the interest for future research in considering alternative low flow timing approaches (e.g., *Van Huijgevoort et al.* [2012]).

The 1988 drought and the 1993 flood events were overall well captured by all the GIMs, with runoff variations compared to the observed data of comparable spatial pattern and intensity. This result provides insightful confidence on the capability of these models to simulate single specific multimonth events on both ends of the runoff spectrum.

To complement our evaluation of the GIMS, we carried out an in-depth pairwise comparison between observations and model outputs using a subset of streamflow gauges and corresponding grid cells. The GIMs' performance was assessed on all hydrological indices through a number of performance metrics. Results from this assessment indicate a better performance of the GIMs in describing the medium flow and its timing compared to the annual maximum and minimum flows. This could be expected as it reflects the increased difficulty of the GIMs in describing extreme events whose occurrence is more erratic (especially high flows) and whose onset is harder to capture (especially low flows) considering the uncertainties that are cascaded across the different model components, and the limited knowledge of the world. In general, the ensemble median proved to perform better and to be more stable than any of the GIMs individually as seen in other previous studies. This is consistent with *Stahl et al*. [2012] who used the same data set over Europe. They found both a better performance of the ensemble mean over each GIM and a decreasing agreement between observed and modeled trends as they moved from annual mean runoff to the tails of the distribution. They also found the widest spread among models for low flow trends, in the same way the performance metrics of our low flow indices were more variable than medium and high flows. *Tallaksen and Stahl* [2014] also revealed considerable model dispersion in simulating temporal and spatial persistence of drought. They warned about the importance of validating GIMs specifically for hydrological drought when analyzing drought characteristics from a limited number of models. Generally, this is valid for all hydrological studies that involve the use of GIMs: the validation of their performance in either high, medium, or low flows is key depending on the flow of interest. However, this is particularly relevant for low flows, because GIMs tend to provide larger uncertainties (i.e., intermodel spread) than the other flow types (high and medium) due to their high sensitivity to model structure and parameterization [*Wang et al*., 2009].

Multimodel studies like WaterMIP comprise many participating GIMs, each of them developed using different conceptual approaches. This make it difficult to identify the reasons for different model behavior and more generally to attribute model error. For instance, conducting parameter sensitivity on an ensemble of GIMs is theoretically possible, but unrealistic in practice, as it would require full control over each model. Similarly, the effect of calibration on model output is rarely quantified for large scale models, which rarely undergo calibration as the traditional catchment models do. The study by *Müller Schmied et al*. [2014] provides some insights in this regard as it uses the only WaterMIP calibrated model WaterGAP in different configurations to investigate the sensitivity of simulated freshwater fluxes and storages to five major sources of uncertainty: climate forcing, land cover input, model structure/refinement, human water use, and calibration against observed mean river discharge. They find that the largest impacts on freshwater fluxes and water storages came from calibration and model structure (e.g., modeling groundwater depletion) and to a lesser extent to alternative climate forcings, and land cover data, whose effects tend to compensate and cancel each other out. In a study on the MacPDM model *Gosling and Arnell* [2011] present a sensitivity analysis and report that simulated runoff is more sensitive to the choice of method to calculate potential evaporation (PE) (having tested Penman-Mointeith and Priestley-Taylor) than to perturbations in soil moisture capacity and field capacity for each specific vegetation type. In particular, they suggest that regional projections from GIMs are likely to be conditional upon the PE method applied, because each method may be more reliable in dry rather than in wet regions. For instance, for much of the United States, the Priestley-Taylor is associated with positive runoff anomalies compared the Penman-Monteith (used in our study), and the situation is reversed for wetter regions. The same authors also report that MacPDM, when running with monthly input data (in our study, however, forcing data from WaterMIP is provided at daily time step), produces a negative runoff bias in several regions of the world where day-to-day variability in relative humidity is high and attribute this bias to difficulties of this GIM in disaggregating monthly relative humidity into daily data.

These results represent a key step toward an improved understanding of the ability of the models to reproduce the hydrologic processes and their temporal changes over the central United States. In particular, this study provides a benchmark for the application of data from intercomparison experiments that make use of this type of GIMs. Building confidence in the models' ability to capture the overall temporal trends and the timing of the hydrology at the regional scale is of great importance for the climate impact studies that will follow, in light of the large socio-economic impacts of too little or too much water will have over this region in a warmer climate.

Acknowledgments
We thank the land-surface and hydrology modeling groups participating to the Watch Project and WaterMIP Project, whose model output was used in this study. To access the data please refer to http://www.eu-watch.org/data_availability for the Watch/WaterMIP modeled data and http://waterdata.usgs.gov/nwis/sw for the streamflow gauges data. Ignazio Giuntoli was funded by a PhD. scholarship from the United Kingdom Natural Environment Research Council (NE/YXS1270382) to the University of Birmingham in collaboration with Centre for Ecology and Hydrology. Gabriele Villarini and Iman Mallakpour acknowledge financial support from the USACE Institute for Water Resources, the Iowa Flood Center, and IIHR-Hydroscience & Engineering. The comments and suggestions by four anonymous reviewers and Van Loon are also acknowledged.

References

Beven, K. (2011), *Rainfall-Runoff Modelling: The Primer*, John Wiley, Hoboken, N. J. [Available at: http://books.google.com/books?hl=en&lr=&id=el-jjlTirlAC&oi=fnd&pg=PR7&dq=Rainfall-runoff+modelling:+the+primer&ots=9RyN7RPj39&sig=vz3TObCO9r6Ld-2UTz6clFGBzAo.]

Christensen, J. H., and O. B. Christensen (2003), Climate modelling: Severe summertime flooding in Europe, *Nature*, *421*(6925), 805–806, doi:10.1038/421805a.

Döll, P., and H. M. Schmied (2012), How is the impact of climate change on river flow regimes related to the impact on mean annual runoff? A global-scale analysis, *Environ. Res. Lett.*, *7*(1), 014037, doi:10.1088/1748-9326/7/1/014037.

Douglas, E. M., R. M. Vogel, and C. N. Kroll (2000), Trends in floods and low flows in the United States: Impact of spatial correlation, *J. Hydrol.*, *240*(1–2), 90–105, doi:10.1016/S0022-1694(00)00336-X.

Giuntoli, I., B. Renard, J.-P. Vidal, and A. Bard (2013), Low flows in France and their relationship to large-scale climate indices, *J. Hydrol.*, *482*, 105–118, doi:10.1016/j.jhydrol.2012.12.038.

Gosling, S. N., and N. W. Arnell (2011), Simulating current global river runoff with a global hydrological model: Model revisions, validation, and sensitivity analysis, *Hydrol. Processes*, *25*(7), 1129–1145, doi:10.1002/hyp.7727.

Gudmundsson, L., T. Wagener, L. M. Tallaksen, and K. Engeland (2012a), Evaluation of nine large-scale hydrological models with respect to the seasonal runoff climatology in Europe, *Water Resour. Res.*, *48*, W11504, doi:10.1029/2011WR010911.

Gudmundsson, L., et al. (2012b), Comparing large-scale hydrological model simulations to observed runoff percentiles in Europe, *J. Hydrometeorol.*, *13*(2), 604–620, doi:10.1175/JHM-D-11-083.1.

Haddeland, I., et al. (2011), Multimodel estimate of the global terrestrial water balance: Setup and first results, *J. Hydrometeorol.*, *12*(5), 869–884, doi:10.1175/2011JHM1324.1.

Hagemann, S., et al. (2013), Climate change impact on available water resources obtained using multiple global climate and hydrology models, *Earth Syst. Dyn.*, *4*(1), 129–144, doi:10.5194/esd-4-129-2013.

Hannah, D. M., S. Demuth, H. A. J. Van Lanen, U. Looser, C. Prudhomme, G. Rees, K. Stahl, and L. M. Tallaksen (2011), Large-scale river flow archives: Importance, current status and future needs, *Hydrol. Processes*, *25*(7), 1191–1200, doi:10.1002/hyp.7794.

Held, I. M., and B. J. Soden (2006), Robust responses of the hydrological cycle to global warming, *J. Clim.*, *19*(21), 5686–5699, doi:10.1175/JCLI3990.1.

Helsel, D. R., and R. M. Hirsch (1992), *Statistical Methods in Water Resources*, vol. 49, Elsevier, Amsterdam. [Available at http://www.scopus.com/scopus/inward/record.url?eid=2-s2.0-0027065640&partnerID=40&rel=R5.5.0.]

Hirsch, R. M., and K. R. Ryberg (2012), Has the magnitude of floods across the USA changed with global CO2 levels?, *Hydrol. Sci. J.*, *57*(1), 1–9, doi:10.1080/02626667.2011.621895.

Hunger, M., and P. Döll (2008), Value of river discharge data for global-scale hydrological modeling, *Hydrol. Earth Syst. Sci.*, *12*, 841–861. [Retrieved from https://hal.archives-ouvertes.fr/hal-00305173/.]

Huntington, T. (2006), Evidence for intensification of the global water cycle: Review and synthesis, *J. Hydrol.*, *319*(1–4), 83–95, doi:10.1016/j.jhydrol.2005.07.003.

Lins, H. F., and J. R. Slack (1999), Streamflow trends in the United States, *Geophys. Res. Lett.*, *26*(2), 227–230, doi:10.1029/1998GL900291.

Lins, H. F., and J. R. Slack (2005), Seasonal and Regional Characteristics of U.S. Streamflow Trends in the United States from 1940 to 1999, *Phys. Geogr.*, *26*(6), 489–501, doi:10.2747/0272-3646.26.6.489.

Mallakpour, I., and G. Villarini (2015), The changing nature of flooding across the central United States, *Nat. Clim. Change*, 1–5, doi:10.1038/nclimate2516.

McCabe, G. J., and D. M. Wolock (2002), A step increase in streamflow in the conterminous United States, *Geophys. Res. Lett.*, *29*(24), 2185, doi:10.1029/2002GL015999.

Milly, P. C. D., R. T. Wetherald, K. A. Dunne, and T. L. Delworth (2002), Increasing risk of great floods in a changing climate, *Nature*, *415*(6871), 514–7, doi:10.1038/415514a.

Milly, P. C. D., K. A. Dunne, and A. V. Vecchia (2005), Global pattern of trends in streamflow and water availability in a changing climate, *Nature*, *438*(7066), 347–50, doi:10.1038/nature04312.

Mölders, N. (2005), Feedbacks at the hydro-meteorological interface, in *Coupled Models for the Hydrological Cycle - Integrating Atmosphere, Biosphere, and Pedosphere*, edited by S. Bronstert et al., pp. 192–208, Springer, Berlin.

Moore, J. N., J. T. Harper, and M. C. Greenwood (2007), Significance of trends toward earlier snowmelt runoff, Columbia and Missouri Basin headwaters, western United States, *Geophys. Res. Lett.*, *34*, L16402, doi:10.1029/2007GL031022.

Müller Schmied, H., S. Eisner, D. Franz, M. Wattenbach, F. T. Portmann, M. Flörke, and P. Döll (2014), Sensitivity of simulated global-scale freshwater fluxes and storages to input data, hydrological model structure, human water use and calibration, *Hydrol. Earth Syst. Sci.*, *18*(9), 3511–3538, doi:10.5194/hess-18-3511-2014.

Peterson, T. C., et al. (2013), Monitoring and understanding changes in heat waves, cold waves, floods, and droughts in the United States: State of knowledge, *Bull. Am. Meteorol. Soc.*, *94*(6), 821–834, doi:10.1175/BAMS-D-12-00066.1.

Prudhomme, C., and H. Davies (2008), Assessing uncertainties in climate change impact analyses on the river flow regimes in the UK. Part 2: Future climate, *Clim. Change*, *93*(1–2), 197–222, doi:10.1007/s10584-008-9461-6.

Prudhomme, C., S. Parry, J. Hannaford, D. B. Clark, S. Hagemann, and F. Voss (2011), How well do large-scale models reproduce regional hydrological extremes in Europe?, *J. Hydrometeorol.*, *12*(6), 1181–1204, doi:10.1175/2011JHM1387.1.

Prudhomme, C., et al. (2014), Hydrological droughts in the 21st century, hotspots and uncertainties from a global multimodel ensemble experiment, *Proc. Natl. Acad. Sci. U.S.A.*, *111*(9), 3262–3267, doi:10.1073/pnas.1222473110.

Rasmussen, T. J., and Perry, C. A. (2001), Trends in peak flows of selected streams in Kansas, *U.S. Geol. Surv. Water Resour. Invest. Rep. 01-4203*. [Retrieved from http://ks.water.usgs.gov/pubs/reports/wrir.01-4203.html#HDR15.]

Smith, M. B., et al. (2012), Results of the DMIP 2 Oklahoma experiments, *J. Hydrol.*, *418-419*, 17–48, doi:10.1016/j.jhydrol.2011.08.056.

Stahl, K., L. M. Tallaksen, J. Hannaford, and H. A. J. Van Lanen (2012), Filling the white space on maps of European runoff trends: Estimates from a multi-model ensemble, *Hydrol. Earth Syst. Sci.*, *16*(7), 2035–2047, doi:10.5194/hess-16-2035-2012.

Stewart, I. T., D. R. Cayan, and M. D. Dettinger (2005), Changes toward earlier streamflow timing across Western North America, *J. Clim.*, *18*(8), 1136–1155, doi:10.1175/JCLI3321.1.

Stott, P. A., N. P. Gillett, G. C. Hegerl, D. J. Karoly, D. A. Stone, X. Zhang, and F. Zwiers (2010), Detection and attribution of climate change: A regional perspective, *Wiley Interdiscip. Rev. Clim. Change*, *1*(2), 192–211, doi:10.1002/wcc.34.

Tallaksen, L. M., and K. Stahl (2014), Spatial and temporal patterns of large-scale droughts in Europe: Model dispersion and performance, *Geophys. Res. Lett.*, *41*, 429–434, doi:10.1002/2013GL058573.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl (2012), An overview of CMIP5 and the experiment design, *Bull. Am. Meteorol. Soc.*, *93*(4), 485–498, doi:10.1175/BAMS-D-11-00094.1.

Uppala, S. M., et al. (2005), The ERA-40 re-analysis, *Q. J. R. Meteorol. Soc.*, *131*(612), 2961–3012, doi:10.1256/qj.04.176.

Van Huijgevoort, M. H. J., P. Hazenberg, H. A. J. Van Lanen, and R. Uijlenhoet (2012), A generic method for hydrological drought identification across different climate regions, *Hydrol. Earth Syst. Sci.*, *16*(8), 2437–2451, doi:10.5194/hess-16-2437-2012.

Van Loon, A. F., M. H. J. Van Huijgevoort, and H. A. J. Van Lanen (2012), Evaluation of drought propagation in an ensemble mean of large-scale hydrological models, *Hydrol. Earth Syst. Sci.*, *16*(11), 4057–4078, doi:10.5194/hess-16-4057-2012.

Villarini, G., J. A. Smith, M. L. Baeck, and W. F. Krajewski (2011), Examining flood frequency distributions in the Midwest U.S, *J. Am. Water Resour. Assoc.*, *47*(3), 447–463, doi:10.1111/j.1752-1688.2011.00540.x.

Villarini, G., J. A. Smith, and G. A. Vecchi (2013), Changing frequency of heavy rainfall over the central United States, *J. Clim.*, *26*(1), 351–357, doi:10.1175/JCLI-D-12-00043.1.

Vogel, R. M., C. Yaindl, and M. Walter (2011), Nonstationarity: Flood magnification and recurrence reduction factors in the United States, *J. Am. Water Resour. Assoc.*, *47*(3), 464–474, doi:10.1111/j.1752-1688.2011.00541.x.

Wang, A., T. J. Bohn, S. P. Mahanama, R. D. Koster, and D. P. Lettenmaier (2009), Multimodel ensemble reconstruction of drought over the continental United States, *J. Clim.*, *22*(10), 2694–2712, doi:10.1175/2008JCLI2586.1.

Warszawski, L., K. Frieler, V. Huber, F. Piontek, O. Serdeczny, and J. Schewe (2014), The Inter-Sectoral Impact Model Intercomparison Project (ISI-MIP): Project framework, *Proc. Natl. Acad. Sci. U.S.A.*, *111*(9), 3228–32, doi:10.1073/pnas.1312330110.

Weedon, G. P., S. Gomes, P. Viterbo, W. J. Shuttleworth, E. Blyth, H. Österle, J. C. Adam, N. Bellouin, O. Boucher, and M. Best (2011), Creation of the WATCH Forcing Data and its use to assess global and regional reference crop evaporation over land during the twentieth century, *J. Hydrometeorol.*, *12*(5), 823–848, doi:10.1175/2011JHM1369.1.

Whitfield, P. H., D. H. Burn, J. Hannaford, H. Higgins, G. A. Hodgkins, T. Marsh, and U. Looser (2012), Reference hydrologic networks I. The status and potential future directions of national reference hydrologic networks for detecting trends, *Hydrol. Sci. J.*, *57*(8), 1562–1579, doi:10.1080/02626667.2012.728706.

Williams, K., and D. B. Clark (2014), *Hadley Centre Technical Note 96 Disaggregation of Daily data in JULES*, pp. 1–28, Exeter, U. K.