GMD Perspective: the quest to improve the evaluation of groundwater representation in continental to global scale models

Tom Gleeson [1,2], Thorsten Wagener [3], Petra Döll [4], Samuel C Zipper [1, 5], Charles West [3], Yoshihide Wada[6], Richard Taylor [7], Bridget Scanlon [8], Rafael Rosolem[3], Shams Rahman[3], Nurudeen Oshinlaja [9], Reed Maxwell [10], Min-Hui Lo [11], Hyungjun Kim [12,13,14], Mary Hill [15], Andreas Hartmann [16,3], Graham Fogg [17], James S. Famiglietti [18], Agnès Ducharne [19], Inge de Graaf [20,21], Mark Cuthbert [11,22], Laura Condon [23], Etienne Bresciani [24], Marc F.P. Bierkens [25, 26]

[1] Department of Civil Engineering, University of Victoria, Canada

[2] School of Earth and Ocean Sciences, University of Victoria, Canada

[3] Department of Civil Engineering, University of Bristol, UK & Cabot Institute, University of Bristol, UK.

[4] Institut für Physische Geographie, Goethe-Universität Frankfurt am Main and Senckenberg Leibniz Biodiversity and Climate Research Centre Frankfurt (SBiK-F), Frankfurt am Main, Germany

[5] Kansas Geological Survey, University of Kansas, USA

[6] International Institute for Applied Systems Analysis, Laxenburg, Austria

[7] Department of Geography, University College London, UK

[8] Bureau of Economic Geology, The University of Texas at Austin, USA

[9] School of Earth and Environmental Sciences & Water Research Institute, Cardiff University, UK

[10] Department of Geology and Geological Engineering, Colorado School of Mines, USA

[11] Department of Atmospheric Sciences, National Taiwan University, Taiwan

[12] Moon Soul Graduate School of Future Strategy, Korea Advanced Institute of Science Technology, Korea

[13] Department of Civil and Environmental Engineering Korea Advanced Institute of Science Technology, Korea

[14] Institute of Industrial Science, The University of Tokyo, Japan

[15] Department of Geology, University of Kansas, USA

[16] Chair of Hydrological Modeling and Water Resources, University of Freiburg, Germany

[17] Department of Land, Air and Water Resources and Earth and Planetary Sciences, University of California, Davis, USA

[18] School of Environment and Sustainability and Global Institute for Water Security, University of Saskatchewan, Saskatoon, Canada

[19] Sorbonne Université, CNRS, EPHE, IPSL, UMR 7619 METIS, Paris, France

[20] Chair or Environmental Hydrological Systems, University of Freiburg, Germany

[21] Water Systems and Global Change Group, Wageningen University, Wageningen, Netherlands

[22] School of Civil and Environmental Engineering, The University of New South Wales, Sydney, Australia

[23] Department of Hydrology & Atmospheric Sciences, University of Arizona, Tucson, Arizona, USA

[24] Center for Advanced Studies in Arid Zones (CEAZA), La Serena, Chile

[25] Physical Geography, Utrecht University, Utrecht, Netherlands

[26] Deltares, Utrecht, Netherlands

Abstract

Continental- to global-scale hydrologic and land surface models increasingly include representations of the groundwater system. Such large-scale models are essential for

examining, communicating, and understanding the dynamic interactions between the Earth System above and below the land surface as well as the opportunities and limits of groundwater resources. We argue that both large-scale and regional-scale groundwater models have utility, strengths and limitations so continued modeling at both scales is essential and mutually beneficial. A crucial quest is how to evaluate the realism, capabilities and performance of large-scale groundwater models given their modeling purpose of addressing large-scale science or sustainability questions as well as limitations in data availability and commensurability. Evaluation should identify if, when or where large-scale models achieve their purpose or where opportunities for improvements exist so that such models better achieve their purpose. We suggest that reproducing the spatio-temporal details of regional-scale models and matching local data is not a relevant goal. Instead, it is important to decide on reasonable model expectations regarding when a large scale model is performing 'well enough' in the context of its specific purpose. The decision of reasonable expectations is necessarily subjective even if the evaluation criteria are quantitative. Our objective is to provide recommendations for improving the evaluation of groundwater representation in continental- to global-scale models. We describe current modeling strategies and evaluation practices, and subsequently discuss the value of three evaluation strategies: 1) comparing model outputs with available observations of groundwater levels or other state or flux variables (observation-based evaluation); 2) comparing several models with each other with or without reference to actual observations (model-based evaluation); and 3) comparing model behavior with expert expectations of hydrologic behaviors in particular regions or at particular times (expert-based evaluation). Based on evolving practices in model evaluation as well as innovations in observations, machine learning and expert elicitation, we argue that combining observation-, model-, and expert-based model evaluation approaches, while accounting for commensurability issues, may significantly improve the realism of groundwater representation in large-scale models. Thus advancing our ability for quantification, understanding, and prediction of crucial Earth science and sustainability problems. We encourage greater community-level communication and cooperation on this quest, including among global hydrology and land surface modelers, local to regional hydrogeologists, and hydrologists focused on model development and evaluation.

1. INTRODUCTION: why and how is groundwater modeled at continental to global scales?

Groundwater is the largest human- and ecosystem-accessible freshwater storage component of the hydrologic cycle (UNESCO, 1978; Margat & Van der Gun, 2013; Gleeson et al., 2016). Therefore, better understanding of groundwater dynamics is critical at a time when the ‘great acceleration’ (Steffen et al., 2015) of many human-induced processes is increasing stress on water resources (Wagener et al., 2010; Montanari et al., 2013; Sivapalan et al., 2014; van Loon et al., 2016), especially in regions with limited data availability and analytical capacity. Groundwater is often considered to be an inherently regional rather than global resource or system. This is partially reasonable because local to regional peculiarities of hydrology, politics and culture are paramount to groundwater resource management (Foster et al. 2013) and groundwater dynamics in different continents are less directly connected and coupled than atmospheric dynamics. Regional-scale analysis and models are essential for addressing local to regional groundwater issues. Generally, regional scale modeling is a mature, well-established field (Hill & Tiedeman, 2007; Kresic, 2009; Zhou & Li, 2011; Hiscock & Bense, 2014; Anderson et al. 2015a) with clear and robust model evaluation guidelines (e.g. ASTM, 2016; Barnett et al., 2012). Regional models have been developed around the world; for example, Rossman & Zlotnik (2014) and Vergnes et al. (2020) synthesize regional-scale groundwater models across the western United States and Europe, respectively.

Yet, important global aspects of groundwater both as a resource and as part of the Earth System are emerging (Gleeson et al. 2020). First, our increasingly globalized world trades virtual groundwater and other groundwater-dependent resources in the food-energy-water nexus, and groundwater often crosses borders in transboundary aquifers. A solely regional approach can be insufficient to analysing and managing these complex global interlinkages. Second, from an Earth system perspective, groundwater is part of the hydrological cycle and connected to the atmosphere, oceans and the deeper lithosphere. A solely regional approach is insufficient to uncover and understand the complex interactions of groundwater within the Earth System and teleconnections, which are groundwater levels or flows in one region linked to geographically separated regions via physical or socio-economic processes. Regional approaches generally focus on important aquifers which underlie only a portion of the world's land mass or population and do not include many other parts of the land surface that may be important for processes like surface water-groundwater exchange flows and evapotranspiration. A global approach is also essential to assess the impact of groundwater depletion on sea level rise, since groundwater storage loss rate on all continents of the Earth must be aggregated. Thus, we argue that groundwater is simultaneously a local, regional, and increasingly global resource and system and that examining groundwater problems, solutions, and interactions at all scales is crucial. As a consequence, we urgently require predictive understanding about how groundwater, used by humans and connected with other components of the Earth System, operates at a variety of scales.

Based on the arguments above for considering global perspectives on groundwater, we see four specific purposes of representing groundwater in continental- to global-scale hydrological or land surface models and their climate modeling frameworks:

(1) To understand and quantify interactions between groundwater and past, present and future climate. Groundwater systems can have far-reaching effects on climate affecting modulation of surface energy and water partitioning with a long-term memory (Anyah et al., 2008; Maxwell and Kollet, 2008; Koirala et al. 2013; Krakauer et al., 2014; Maxwell et al., 2016; Taylor, et al., 2013a; Meixner et et, 2018; Wang et al., 2018; Keune et al., 2018). While there have been significant advances in understanding the role of lateral groundwater flow on evapotranspiration (Maxwell & Condon, 2016; Bresciani et al, 2016), the interactions between climate and groundwater over longer time scales (Cuthbert et al., 2019) as well as between irrigation, groundwater, and climate (Condon and Maxwell, 2019; Condon et al 2020) remain largely unresolved. Additionally, it is well established that old groundwater with slow turnover times are common at depth (Befus et al. 2017; Jasechko et al. 2017). Groundwater connections to

the atmosphere are well documented in modeling studies (e.g. Forrester and Maxwell, 2020). Previous studies have demonstrated connections between the atmospheric boundary layer and water table depth (e.g. Maxwell et al 2007; Rahman et al, 2015), under land cover disturbance (e.g. Forrester et al 2018), under extremes (e.g. Kuene et al 2016) and due to groundwater pumping (Gilbert et al 2017). While a number of open source platforms have been developed to study these connections (e.g. Maxwell et al 2011; Shrestha et al 2014; Sulis, 2017), these platforms are regional to continental in extent. Recent work has shown global impacts of groundwater on atmospheric circulation (Wang et al 2018), but groundwater is still quite simplified in this study.

(2) To understand and quantify two-way interactions between groundwater, the rest of the hydrologic cycle, and the broader Earth System. As the main storage component of the freshwater hydrologic cycle, groundwater systems support baseflow levels in streams and rivers, and thereby ecosystems and agricultural productivity and other ecosystem services in both irrigated and rainfed systems (Scanlon et al., 2012; Qiu et al., 2019; Visser, 1959; Zipper et al., 2015, 2017). When pumped groundwater is transferred to oceans (Konikow 2011; Wada et al., 2012; Döll et al., 2014a; Wada, 2016; Caceres et al., 2020; Luijendijk et al. 2020), resulting sea-level rise can impact salinity levels in coastal aquifers, and freshwater and solute inputs to the ocean (Moore, 2010; Sawyer et al., 2016). Difficulties are complicated by international trade of virtual groundwater which causes aquifer stress in disparate regions (Dalin et al., 2017)

(3) To inform water decisions and policy for large, often transboundary groundwater systems in an increasingly globalized world (Wada & Heinrich, 2013; Herbert & Döll, 2019). For instance, groundwater recharge from large-scale models has been used to quantify groundwater resources in Africa, even though large-scale models do not yet include all recharge processes that are important in this region (Taylor et al., 2013b; Jasechko et al. 2014; Cuthbert et al., 2019; Hartmann et al., 2017).

(4) To create visualizations and interactive opportunities that inform citizens and consumers, whose decisions have global-scale impacts, about the state of groundwater all around the world such as the World Resources Institute’s Aqueduct website (https://www.wri.org/aqueduct), a decision-support tool to identify and evaluate global water risks.

The first two purposes are science-focused while the latter two are sustainability-focused. In sum, continental- to global-scale hydrologic models incorporating groundwater offer a coherent scientific framework to examine the dynamic interactions between the Earth System above and below the land surface, and are compelling tools for conveying the opportunities and limits of groundwater resources to people so that they can better manage the regions they live in, and better understand the world around them. We consider both large-scale and regional-scale models to be useful practices that should both continue to be conducted rather than one replacing another. Ideally large-scale and regional-scale models should benefit from the other since each has strengths and weaknesses and together the two practices enrich our understanding and support the management of groundwater across scales (Section 2).

The challenge of incorporating groundwater processes into continental- or global-scale models is formidable and sometimes controversial. Some of the controversy stems from unanswered questions about how best to represent groundwater in the models whereas some comes from skepticism about the feasibility of modelling groundwater at non-traditional scales. We advocate for the representation of groundwater stores and fluxes in continental to global models for the four reasons described above. We do not claim to have all the answers on how best to meet this challenge. We contend, however, that the hydrologic community needs to work deliberately and constructively towards effective representations of groundwater in global models.

Driven by the increasing recognition of the purpose of representing groundwater in continental- to global-scale models, many global hydrological models and land surface models have incorporated groundwater to varying levels of complexity depending on the model provenance and purpose. Different from regional-scale groundwater models that generally focus on subsurface dynamics, the focus of these models is on estimating either runoff and streamflow (hydrological models) or land-atmosphere water and energy exchange (land surface models). Simulation of groundwater storages and hydraulic heads mainly serve to quantify baseflow that affects streamflow during low flow periods or capillary rise that increases evapotranspiration. Some land-surface models use approaches based on the topographic index to simulate fast surface and slow subsurface runoff based on the fraction of saturated area in the grid cell (Clark et al., 2015; Fan et al., 2019); groundwater in these models does not explicitly have water storage or hydraulic heads (Famiglietti & Wood, 1994; Koster et al., 2000; Niu et al., 2003; Takata et al., 2003). In many hydrological models, groundwater is represented as a linear reservoir that is fed by groundwater recharge and drains to a river in the same grid cell (Müller Schmied et al., 2014; Gascoin et al., 2009; Ngo-Duc et al., 2007). Time series of groundwater storage but not hydraulic heads are computed. This prevents simulation of lateral groundwater flow between grid cells, capillary rise and two-way exchange flows between surface water bodies and groundwater (Döll et al., 2016). However, representing groundwater as a water storage compartment that is connected to soil and surface water bodies by groundwater recharge and baseflow and is affected by groundwater abstractions and returns, enables global-scale assessment of groundwater resources and stress (Herbert and Döll, 2019) and groundwater depletion (Döll et al., 2014a; Wada et al., 2014; de Graaf et al., 2014). In some land surface models, the location of the groundwater table with respect to the land surface is simulated within each grid cell to enable simulation of capillary rise (Niu et al., 2007) but, as in the case of simulating groundwater as a linear reservoir, lateral groundwater transport or two-way surface water-groundwater exchange cannot be simulated with this approach.

Increasingly, models for simulating groundwater flows between all model grid cells in entire countries or globally have been developed, either as stand-alone models or as part of hydrological models (Vergnes & Decharme, 2012; Fan et al., 2013; Lemieux et al. 2008; de Graaf et al., 2017; Kollet et al., 2017; Maxwell et al., 2015; Reinecke et al., 2018, de Graaf et al 2019). The simulation of groundwater in large-scale models is a nascent and rapidly developing field with significant computational and parameterization challenges which have led to significant and important efforts to develop and evaluate individual models. It is important to note that herein 'large-scale models' refer to models that are laterally extensive across multiple regions (hundreds to thousands of kilometers) and generally include the upper tens to hundreds of meters of subsurface and have resolutions sometimes as small as ~1 km. In contrast, 'regional-scale' models (tens to hundreds of kilometers) have long been developed for a specific region or aquifer and can include greater depths and resolutions, more complex hydrostratigraphy and are often developed from conceptual models with significant regional knowledge. Regional-scale models include a diverse range of approaches from stand-alone groundwater models (i.e., representing surface water and vadose zone processes using boundary conditions such as recharge) to fully integrated groundwater-surface water models. In the future, large-scale models could be developed in a number of different directions which we only briefly introduce here to maintain our primary focus on model evaluation. One important direction is clearer representation of three-dimensional geology and heterogeneity including karst (Condon et al. in review) which should be considered as part of conceptual model development prior to numerical model implementation.

Now that a number of models that represent groundwater at continental to global scales have been developed and will continue evolving, it is equally important that we advance how we evaluate these models. To date, large-scale model evaluation has largely focused on individual models, with inconsistent practices between models and little community-level discussion or cooperation, that lack the rigor of regional-scale model evaluation. Overall, we have only a partial and piecemeal understanding of the capabilities and limitations of different approaches to representing groundwater in large-scale models. Our objective is to provide clear recommendations for evaluating groundwater representation in continental and global models**.** We focus on model evaluation because this is the heart of model trust and reproducibility (Hutton et al., 2016) and improved model evaluation will guide how and where it is most important to focus future model development. We describe current model evaluation practices (Section 2) and consider diverse and uncertain sources of information, including observations, models, and experts to holistically evaluate the simulation of groundwater-related fluxes, stores and hydraulic heads (Section 3). We stress the need for an iterative and open-ended process of model improvement through continuous model evaluation against the different sources of information. We explicitly contrast the terminology used herein of 'evaluation' and 'comparison' against terminology such as 'calibration' or 'validation' or 'benchmarking', which suggests a modelling process that is at some point complete. We extend previous commentaries advocating improved hydrologic process representation and evaluation in large-scale hydrologic models (Clark et al. 2015; Melsen et al. 2016) by adding expert-elicitation and machine learning for more holistic evaluation. We also consider model objective and model evaluation across the diverse hydrologic landscapes which can both uncover blindspots in model development. It is important to note that we do not consider water quality or contamination, even though water quality or contamination is important for water resources, management and sustainability, since large-scale water quality models are in their infancy (van Vliet et al., 2019)

We bring together somewhat disparate scientific communities as a step towards greater community-level cooperation on these challenges, including global hydrology and land surface modelers, local to regional hydrogeologists, and hydrologists focused on model development and evaluation. We see three audiences beyond those currently directly involved in large-scale groundwater modeling that we seek to engage to accelerate model evaluation: 1) regional hydrogeologists who could be reticent about global models, and yet have crucial knowledge and data that would improve evaluation; 2) data scientists with expertise in machine learning, artificial intelligence etc. whose methods could be useful in a myriad of ways; and 3) the multiple Earth Science communities that are currently working towards integrating groundwater into a diverse range of models so that improved evaluation approaches are built directly into model development.

2. CURRENT MODEL EVALUATION PRACTICES

Here we provide a brief overview of current large-scale groundwater models, the synergies and differences between regional-scale and large-scale model evaluation and development as well as the imitations of current evaluation practices for large-scale models.

2.1 Brief overview of current large-scale groundwater models

Various large-scale models exist along a spectrum of model complexity, which can make it difficult to determine the most appropriate model for a specific application. We developed a simple but systematic classification of current large-scale groundwater models (Table 1) to summarize the main characteristics of existing models for the interdisciplinary audience of GMD. This classification builds on other reviews (Bierkens 2015; Condon et al., in review) and is not exhaustive, nor is it the only way to classify large-scale groundwater models. It is meant to be a first classification attempt that should evolve with time. We suggest that groundwater in current large-scale models can be classified functionally by two aspects that are crucial to how groundwater impacts water, energy, and nutrient budgets. First, whether lateral subsurface flow to a river is simulated within each cell independently of other cells, as 2D lateral groundwater flow between all cells or as 3D groundwater flow. Second, we distinguish two types of coupling between groundwater and related compartments (variably saturated soil zone, surface water, atmospheric processes): ‘one-way’ coupling (for example, recharge is imposed from the surface with no feedback from capillary rise or vegetation uptake, or groundwater flow to the surface does not depend on surface head) from ‘two-way’ coupling involves feedback loops. We also note atmospheric coupling which involves coupling a groundwater-surface model with an atmospheric model to propagate the influence of groundwater from the surface to the atmosphere, and the resulting feedback onto the surface and groundwater. This classification scheme (which could also be called a model typology) is based on a number of model characteristics such as the fluxes, stores and other features (Table 1).

2.2 Synergies between regional-scale and large-scales

Regional-scale and large-scale groundwater models are both governed by the same physical equations and share many of the same challenges. Like large-scale models, some regional-scale models have challenges with representing important regional hydrologic processes such as mountain block recharge (Markovich et al. 2019), and data availability challenges (such as the lack of reliable subsurface parameterization and hydrologic monitoring data) are common. We propose there are largely untapped potential synergies between regional-scale and large-scale models based on these commonalities and the inherent strengths and limitations of each scale (Section 1).

Much can be learned from regional-scale models to inform the development and evaluation of large-scale groundwater models. Regional-scale models are evaluated using a variety of data types, some of which are available and already used at the global scale and some of which are not. In general, the most common data types used for regional-scale groundwater model evaluation match global-scale groundwater models: hydraulic head and either total streamflow or baseflow estimated using hydrograph separation approaches (eg. RRCA, 2003; Woolfenden and Nishikawa, 2014; Tolley et al., 2019). However, numerous data sources unavailable or not currently used at the global scale have also been applied in regional-scale models, such as elevation of surface water features (Hay et al., 2018), existing maps of the potentiometric surface (Meriano and Eyles, 2003), and dendrochronology (Schilling et al., 2014) and stable and radiogenic isotopes for determining water sources and residence times (Sanford, 2011). These and other 'non-classical' observations (Schilling et al. 2019) could be the inspiration for model evaluation of large-scale models in the future but are beyond our scope to discuss. Further, given the smaller domain size of regional-scale models, expert knowledge and local ancillary data sources can be more directly integrated and automated parameter estimation approaches such as PEST are tractable (Leaf et al., 2015; Hunt et al., 2013). We directly build upon this practice of integration of expert knowledge below in Section 3.3.

We propose that there may also be potential benefits of large-scale models for the development of regional-scale models. For instance, the boundary conditions of some regional-scale models could be improved with large-scale model results. The boundary conditions of regional-scale models are often assumed, calibrated or derived from other models or data. In a regional-scale model, increasing the model domain (moving the boundary conditions away from region of interests) or incorporating more hydrologic processes (for example, moving the boundary condition from recharge to the land surface incorporating evapotranspiration and infiltration) both can reduce the impact of boundary conditions on the region and problem of interest. Another potential benefit of large-scale models for regional-scale models is fuller inclusion of large-scale hydrologic and human processes that could further enhance the ability of regional-scale models to address both the science-focused and sustainability-focused purposes described in Section 1. For example, the stronger representation of large-scale atmospheric processes means that the downwind impact of groundwater irrigation on evapotranspiration on precipitation and streamflow can be assessed (DeAngelis et al., 2010; Kustu et al., 2011). Or, the effects of climate change and increased water use that affect the inflow of rivers into the regional modelling domain can be taken from global scale analyses (Wada and Bierkens, 2014). Also, regional groundwater depletion might be largely driven by virtual water trade which can be better represented in global analysis and models than regional-scale models (Dalin et al. 2017). Therefore the processes and results of large-scale models could be used to make regional-scale models even more robust and better address key science and sustainability questions.

Given the strengths of regional models, a potential alternative to development of large-scale groundwater models would be combining or aggregating multiple regional models in a patchwork approach (as in Zell and Sanford, 2020) to provide global coverage. This would have the advantage of better respecting regional differences but potentially create additional challenges because the regional models would have different conceptual models, governing equations, boundary conditions etc. in different regions. Some challenges of this patchwork approach include 1) the required collaboration of a large number of experts from all over the world over a long period of time; 2) regional groundwater flow models alone are not sufficient, they need to be integrated into a hydrological model so that groundwater-soil water and the surface water-groundwater interactions can be simulated; 3) the extent of regional aquifers does not necessarily coincide with the extent of river basins; and 4) the bias of regional groundwater models towards important aquifers which as described above, underlie only a portion of the world's land mass or population and may bias estimates of fluxes such as surface water-groundwater exchange or evapotranspiration. Given these challenges, we argue that a patchwork approach of integrating multiple regional models is a compelling idea but likely insufficient to achieve the purposes of large-scale groundwater modeling described in Section 1. Although this nascent idea of aggregating regional models is beyond the scope of this manuscript, we consider this an important future research avenue, and encourage further exploration and improvement of regional-scale model integration from the groundwater modeling community.

2.3 Differences between regional-scale and large-scales

Although there are important similarities and potential synergies across scales, it is important to consider how or if large-scale models are fundamentally different to regional-scale models, especially in ways that could impact evaluation. The primary differences between large-scale and regional-scale models are that large-scale models (by definition) cover larger areas and, as a result, typically include more data-poor areas and are generally built at coarser resolution. These differences impact evaluations in at least five relevant ways:

1) <u>Commensurability errors</u> (also called 'representativeness' errors) occur either when modelled grid values are interpolated and compared to an observation 'point' or when aggregation of observed 'point' values are compared to a modelled grid value (Beven, 2005; Tustison et al., 2001; Beven, 2016; Pappenberger et al., 2009; Rajabi et al., 2018). For groundwater models in particular, commensurability error will depend on the number and locations of observation points, the variability structure of the variables being

compared such as hydraulic head and the interpolation or aggregation scheme applied (Tustison et al., 2001; Pappenberger et al., 2009; Reinecke et al., 2020). Commensurability is a problem for most scales of modelling, but likely more significant the coarser the model. Regional-scale groundwater models typically have fewer (though not insignificant) commensurability issues due to smaller grid cell sizes compared to large-scale models.

2) <u>Specificity to region, objective and model evaluation criteria</u> because regional-scale models are developed specifically for a certain region and modeling or management objective whereas large-scale models are often more general and include different regions. As a result, large-scale models often have greater heterogeneity of processes and parameters, may not adopt the same calibration targets and variables, and are not subject to the policy or litigation that sometimes drives model evaluation of regional-scale models.

3) <u>Computational requirements</u> can be immense for large-scale models which leads to challenges with uncertainty and sensitivity analysis. While some regional-scale models also have large computational demands, large-scale models cover larger domains and are therefore more vulnerable to this potential constraint.

4) <u>Data availability</u> for large-scale models can be limited because they typically include data-poor areas, which leads to challenges when only using observations for model evaluation. While data availability also affects regional-scale models, they are often developed for regions with known hydrological challenges based on existing data and/or modeling efforts are preceded by significant regional data collection from detailed sources (such as

local geological reports) that are not often included in continental to global datasets used for large-scale model parameterization.

5) Subsurface detail in regional-scale models routinely include heterogeneous and anisotropic parameterizations which could be improved in future large-scale models. For example, intense vertical anisotropy routinely induces vertical flow dynamics from vertical head gradients that are tens to thousands of times greater than horizontal gradients which profoundly alter the meaning of the deep and shallow groundwater levels, with only the latter remotely resembling the actual water table. In contrast, currently most large-scale models use a single vertically homogeneous value for each grid cell, or at best have two layers (de Graaf et al,. 2017)

2.4 Limitations of current evaluation practices for large-scale models

Evaluation of large-scale models has often focused on streamflow or evapotranspiration observations but joint evaluation together with groundwater-specific variables is appropriate and necessary (e.g. Maxwell et al. 2015; Maxwell and Condon, 2016). Groundwater-specific variables useful for evaluating the groundwater component of large-scale models include: a) hydraulic head or water table depth; b) groundwater storage and groundwater storage changes which refer to long-term, negative or positive trends in groundwater storage where long-term, negative trends are called groundwater depletion; c) groundwater recharge; d) flows between groundwater and surface water bodies; and e) human groundwater abstractions and return flows to groundwater. It is important to note that groundwater and surface water hydrology communities often have slightly different definitions of terms like recharge and baseflow (Barthel, 2014); we therefore suggest trying to precisely define the meanings of such words using the actual hydrologic fluxes which we do below. Table 2 shows the availability of observational data for these variables but does not evaluate the quality and robustness of observations. Overall there are significant inherent challenges of commensurability and measurability of groundwater observations in the evaluation of large-scale models. We describe the current model evaluation practices for each of these variables here:

a) Simulated hydraulic heads or water table depth in large scale models are frequently compared to well observations, which are often considered the crucial data for groundwater model evaluation. Hydraulic head observations from a large number groundwater wells (>1 million) have been used to evaluate the spatial distribution of steady-state heads (Fan et al., 2013, de Graaf et al., 2015; Maxwell et al., 2015; Reinecke et al., 2019a, 2020). Transient hydraulic heads with seasonal amplitudes (de Graaf et al. 2017), declining heads in aquifers with groundwater depletion (de Graaf et al. 2019) and daily transient heads (Tran et al 2020) have also been compared to well observations. All evaluation with well observations is severely hampered by the incommensurability of point values of observed head with simulated heads that represent averages over cells of a size of tens to hundreds square kilometers; within such a large cell, land surface elevation, which strongly governs hydraulic head, may vary a few hundred meters, and average observed head strongly depends on the number and location of well within the cell (Reinecke et al., 2020). Additional concerns with head observations are the 1) strong sampling bias of wells towards accessible locations, low elevations, shallow water tables, and more transmissive aquifers in wealthy, generally temperate countries (Fan et al., 2019); 2) the impacts of pumping which may or may not be well known; 3) observational errors and uncertainty (Post and von Asmuth, 2013; Fan et al., 2019); and 4) that heads can reflect the poro-elastic effects of mass loading and unloading rather than necessarily aquifer recharge and drainage (Burgess et al, 2017). To date, simulated hydraulic heads have more often been compared to observed heads (rather than water table depth) which results in lower relative errors (Reinecke et al., 2020) because the range of heads (10s to 1000s m head) is much larger than the range of water table depths (<1 m to 100s m).

b) Simulated groundwater storage trends or anomalies in large-scale hydrological models have been evaluated using observations of groundwater well levels combined with estimates of storage parameters, such as specific yield; local-scale groundwater modeling; and translation of regional total water storage trends and anomalies from satellite gravimetry (GRACE: Gravity Recovery And Climate Experiment) to groundwater storage changes by estimating changes in other hydrological storages (Döll et al., 2012; 2014a). Groundwater storage changes volumes and rates have been calculated for numerous aquifers, primarily in the United States, using calibrated groundwater models, analytical approaches, or volumetric budget analyses (Konikow, 2010). Regional-scale models have also been used to simulate groundwater storage trends untangling the impacts of water management during drought (Thatch et al. 2020). Satellite gravimetry (GRACE) is important but has limitations (Alley and Konikow, 2015). First, monthly time series of very coarse-resolution groundwater storage are indirectly estimated from observations of total water storage anomalies by satellite gravimetry (GRACE) but only after model- or observation-based subtraction of water storage changes in glaciers, snow, soil and surface water bodies (Lo et al., 2016; Rodell et al., 2009; Wada, 2016). As soil moisture, river or snow dynamics often dominate total water storage dynamics, the derived groundwater storage dynamics can be so uncertain that severe groundwater drought cannot be detected in this way (Van Loon et al., 2017). Second, GRACE cannot detect the impact of groundwater abstractions on groundwater storage unless groundwater depletion occurs (Döll et al., 2014a,b). Third, the very coarse resolution can lead to incommensurability but in the opposite direction of well observations. It is important to note that the focus is on storage trends or anomalies since total groundwater storage to a specific depth (Gleeson et al., 2016) or in an aquifer (Konikow, 2010) can be estimated but the total groundwater storage in a specific region or cell cannot be simulated or observed unless the depth of interest is specified (Condon et al., 2020).

c) Simulated large-scale groundwater recharge (vertical flux across the water table) has been evaluated using compilations of point estimates of groundwater recharge, results of regional-scale models, baseflow indices, and expert opinion (Döll and Fiedler, 2008; Hartmann et al., 2015) or compared between models (e.g. Wada et al. 2010). In general, groundwater recharge is not directly measurable except by meter-scale lysimeters (Scanlon et al., 2002), and many groundwater recharge methods such as water table fluctuations and chloride mass balance also suffer from similar commensurability issues as water table depth data. Although sometimes an input or boundary condition to regional-scale models, recharge in many large-scale groundwater models is simulated and thus can be evaluated.

d) The flows between groundwater and surface water bodies (rivers, lakes, wetlands) are simulated by many models but are generally not evaluated directly against observations of such flows since they are very rare and challenging. Baseflow (the slowly varying portion of streamflow originating from groundwater or other delayed sources) or streamflow 'low flows' (when groundwater or other delayed sources predominate), generally cannot be used to directly quantify the flows between groundwater and surface water bodies at large scales. Groundwater discharge to rivers can be estimated from streamflow observations only in the very dense gauge network and/or if streamflow during low flow periods is mainly caused by groundwater discharge and not by water storage in upstream lakes, reservoirs or wetlands. These conditions are rarely met in case of streamflow gauges with large upstream areas that can be used for comparison to large-scale model output. de Graaf et al. (2019) compared the simulated timing of changes in groundwater discharge to observations and regional-scale models, but only compared the fluxes directly between the global- and regional-scale models. Due to the challenges of directly observing the flows between groundwater and surface water bodies at large scales, this is not included in the available data in Table 2; instead in Section 3 we highlight the potential for using baseflow or the spatial distribution of perennial, intermittent and ephemeral streams in the future.

e) Groundwater abstractions have been evaluated by comparison to national, state and county scale statistics in the U.S. (Wada et al. 2010, Döll et al., 2012, 2014a, de Graaf et al. 2014). Irrigation is the dominant groundwater use sector in many regions; however, irrigation pumpage is generally estimated from crop water demand and rarely metered. GRACE and other remote sensing data have been used to estimate the irrigation water abstractions (Anderson et al. 2015b). The lack of records or observations of abstraction introduces significant uncertainties into large-scale models and is simulated and thus can be evaluated. Human groundwater abstractions and return flows as well as groundwater recharge and the flows between groundwater and surface water bodies are necessary to simulate storage trends (described above). But each of these are considered separate observations since they each have different data sources and assumptions. Groundwater abstraction data at the well scale are severely hampered by the incommensurability like hydraulic head and recharge described above.

3. HOW TO IMPROVE THE EVALUATION OF LARGE-SCALE GROUNDWATER MODELS

Based on Section 2, we argue that the current model evaluation practices are insufficient to robustly evaluate large-scale models. We therefore propose evaluating large-scale models using at least three strategies (pie-shapes in Figure 1): observation-, model-, and expert-driven evaluation which are potentially mutually beneficial because each strategy has its strengths and weaknesses. We are not proposing a brand new evaluation method here but rather separating strategies to consider the problem of large-scale model evaluation from different but highly interconnected perspectives. All three strategies work together for the common goal of 'improved model large-scale model evaluation' which is what is the centre of Figure 1.

When evaluating large-scale models, it is necessary to first consider reasonable expectations or how to know a model is 'well enough'. Reasonable expectations should be based on the modeling purpose, hydrologic process understanding and the plausibly achievable degree of model realism. First, model evaluation should be clearly linked to the four science- or sustainability-focused purposes of representing groundwater in large-scale models (Section 1) and second, to our understanding of relevant hydrologic processes. The objective of large-scale models cannot be to reproduce the spatio-temporal details that regional-scale models can reproduce. Determining the reasonable expectations is necessarily subjective, but can be approached using observation-, model-, and expert-driven evaluation. As a simple first step in setting realistic expectations, we propose that three physical variables can be used to form more convincing arguments that a large-scale model is well enough: change in groundwater storage, water table depth, and regional fluxes between groundwater and surface water. Below we explore in more detail additional variables and approaches that can support this simple approach.

Across all three model evaluation strategies of observation-, model-, and expert-driven evaluation, we advocate three principles underpinning model evaluation (base of Figure 1), none of which we are the first to suggest but we highlight here as a reminder: 1) model objectives, such as the groundwater science or groundwater sustainability objective summarised in Section 1, are important to model evaluation because they provide the context through which relevance of the evaluation outcome is set; 2) all sources of information (observations, models and experts) are uncertain and this uncertainty needs to be quantified for robust evaluation; and 3) regional differences are likely important for large-scale model evaluation - understanding these differences is crucial for the transferability of evaluation outcomes to other places or times.

We stress that we see the consideration and quantification of uncertainty as an essential need across all three types of model evaluation we describe below, so we discuss it here rather than with model-driven model evaluation (Section 3.2) where uncertainty analysis more narrowly defined would often be discussed. We further note that large-scale models have only been assessed to a very limited degree with respect to understanding, quantifying, and attributing relevant uncertainties. Expanding computing power, developing computationally frugal methods for sensitivity and uncertainty analysis, and potentially employing surrogate models can enable more robust sensitivity and uncertainty analysis such as used in regional-scale models (Habets et al., 2013; Hill, 2006; Hill & Tiedeman, 2007; Reinecke et al., 2019b). For now, we suggest applying computationally frugal methods such as the elementary effect test or local sensitivity analysis (Hill, 2006; Morris, 1991; Saltelli et al., 2000). Such sensitivity and uncertainty analyses should be applied not only to model parameters and forcings but also to model structural properties (e.g. boundary conditions, grid resolution, process simplification, etc.) (Wagener and Pianosi, 2019). This implies that the (independent) quantification of uncertainty in all model elements (observations, parameters, states, etc.) needs to be improved and better captured in available metadata.

We advocate for considering regional differences more explicitly in model evaluation since likely no single model will perform consistently across the diverse hydrologic landscapes of the world (Van Werkhoven et al., 2008). Considering regional differences in large-scale model evaluation is motivated by recent model evaluation results and is already starting to be practiced. Two recent sensitivity analyses of large-scale models reveal how sensitivities to input parameters vary in different regions for both hydraulic heads and flows between groundwater and surface water (de Graaf et al. 2019; Reinecke et al., 2020). In mountain regions, large-scale models tend to underestimate steady-state hydraulic head, possibly due to over-estimated hydraulic conductivity in these regions, which highlights that model performance varies in different hydrologic landscapes. (de Graaf et al., 2015; Reinecke et al. 2019b). Additionally, there are significant regional differences in performance with low flows for a number of large-scale models (Zaherpour et al. 2018) likely because of diverse implementations of groundwater and baseflow schemes. Large-scale model evaluation practice is starting to shift towards highlighting regional differences as exemplified by two different studies that explicitly mapped hydrologic landscapes to enable clearer understanding of regional differences. Reinecke et al. (2019b) identified global hydrological response units which highlighted the spatially distributed parameter sensitivities in a computationally expensive model, whereas Hartmann et al. (2017) developed and evaluated models for karst aquifers in different hydrologic landscapes based on different a priori system conceptualizations. Considering regional differences in model evaluation suggests that global models could in the future consider a patchwork approach of different conceptual models, governing equations, boundary conditions etc. in different regions. Although beyond the scope of this manuscript, we consider this an important future research avenue.

3.1 Observation-based model evaluation

Observation-based model evaluation is the focus of most current efforts and is important because we want models to be consistent with real-world observations. Section 2 and Table 2 highlight both the strengths and limitations of current practices using observations. Despite existing challenges, we foresee significant opportunities for observation-based model evaluation and do not see data scarcity as a reason to exclude groundwater in large-scale models or to avoid evaluating these models. It is important to note that most so-called ‘observations’ are modeled or derived quantities, and often at the wrong scale for evaluating large-scale models (Table 2; Beven, 2019). Given the inherent challenges of direct measurement of groundwater fluxes and stores especially at large scales, herein we consider the word ‘observation’ loosely as any measurements of physical stores or fluxes that are combined with or filtered through models for an output. For example, GRACE gravity measurements are combined with model-based estimates of water storage changes in glaciers, snow, soil and surface water for ‘groundwater storage change observations’ or streamflow measurements are filtered through baseflow separation algorithms for ‘baseflow observations’. The strengths and limitations as well as the data availability and spatial and temporal attributes of different observations are summarized in Table 2 which we hope will spur more systematic and comprehensive use of observations.

Here we highlight nine important future priorities for improving evaluation using available observations. The first five priorities focus on current observations (Table 2) whereas the latter four focus on new methods or approaches:

1) Focus on transient observations of the water table depth rather than hydraulic head observations that are long-term averages or individual times (often following well drilling). Water table depth are likely more robust evaluation metrics than hydraulic head because water table depth reveals great discrepancies and is a complex function of the relationship between hydraulic head and topography that is crucial to predicting system fluxes (including evapotranspiration and baseflow). Comparing transient observations and simulations instead of long-term averages or individual times incorporates more system dynamics of storage and boundary conditions as temporal patterns are more important than absolute values (Heudorfer et al. 2019). For regions with significant groundwater depletion, comparing to declining water tables is a useful strategy (de Graaf et al. 2019), whereas in aquifers without groundwater depletion, seasonally varying water table depths are likely more useful observations (de Graaf et al. 2017).

2) Use baseflow, the slowly varying portion of streamflow originating from groundwater or other delayed sources. Döll and Fiedler (2008) included the baseflow index in evaluating recharge and baseflow has been used to calibrate the groundwater component of a land surface model (Lo et al. 2008, 2010). But the baseflow index (BFI), linear and nonlinear baseflow recession behavior or baseflow fraction (Gnann et al., 2019) have not been used to evaluate any large-scale model that simulates groundwater flows between all model grid cells. There are limitations of using BFI and baseflow

recession characteristics to evaluate large-scale models (Table 2). Using baseflow only makes sense when the baseflow separation algorithm is better than the large-scale model itself, which may not be the case for some large-scale models and only in time periods that can be assumed to be dominated by groundwater discharge. Similarly, using recession characteristics is dependent on an appropriate choice of recession extraction methods. But this remains available and obvious data derived from streamflow or spring flow observations that has been under-used to date.

3) Use the spatial distribution of perennial, intermittent, and ephemeral streams as an observation, which to our best knowledge has not been done by any large-scale model evaluation. The transition between perennial and ephemeral streams is an important system characteristic in groundwater-surface water interactions (Winter et al. 1998), so we suggest that this might be a revealing evaluation criteria although there are similar limitations to using baseflow. The results of both quantifying baseflow and mapping perennial streams depend on the methods applied, they are not useful for quantifying groundwater-surface water interactions when there is upstream surface water storage, and they do not directly provide information about fluxes between groundwater and surface water.

4) Use data on land subsidence to infer head declines or aquifer properties for regions where groundwater depletion is the main cause of compaction

(Bierkens and Wada, 2019). Lately, remote sensing methods such as GPS, airborne and space borne radar and lidar are frequently used to infer land subsidence rates (Erban et al., 2014). Also, a number of studies combine geomechanical modelling (Ortega-Guerrero et al 1999; Minderhoud et al 2017) and geodetic data to explain the main drivers of land subsidence. A few papers (e.g. Zhang and Burbey 2016) use a geomechanical model together with a withdrawal data and geodetic observations to estimate hydraulic and geomechanical subsoil properties.

5) Consider using socio-economic data for improving model input. For example, reported crop yields in areas with predominant groundwater irrigation could be used to evaluate groundwater abstraction rates. Or using well depth data (Perrone and Jasechko, 2019) to assess minimum aquifer depths or in coastal regions and deltas, the presence of deeper fresh groundwater under semi-confining layers.

6) Derive additional new datasets using meta-analysis and/or geospatial analysis such as gaining or losing stream reaches (e.g., from interpolated head measurements close to the streams), springs and groundwater-dependent surface water bodies, or tracers. Each of these new data sources could in principle be developed from available data using methods already applied at regional scales but do not currently have an ‘off the shelf’ global dataset. For example, some large-scale models have been explicitly

compared with residence time and tracer data (Maxwell et al., 2016) which have also been recently compiled globally (Gleeson et al., 2016; Jasechko et al., 2017). This could be an important evaluation tool for large-scale models that are capable of simulating flow paths, or can be modified to do, though a challenge of this approach is the conservativity of tracers. Future meta-analyses data compilations should report on the quality of the data and include possible uncertainty ranges as well as the mean estimates.

7) Use machine learning to identify process representations (e.g. Beven, 2020) or spatiotemporal patterns, for example of perennial streams, water table depths or baseflow fluxes, which might not be obvious in multi-dimensional datasets and could be useful in evaluation. For example, Yang et al. (2019) predicted the state of losing and gaining streams in New Zealand using Random Forest algorithms. A staggering variety of machine learning tools are available and their use is nascent yet rapidly expanding in geoscience and hydrology (Reichstein et al., 2019; Shen, 2018; Shen et al., 2018; Wagener et al., 2020). While large-scale groundwater models are often considered 'data-poor', it may seem strange to propose using data-intensive machine learning methods to improve model evaluation. But some of the data sources are large (e.g over 2 million water level measurements in Fan et al. 2013 although biased in distribution) whereas other observations such as evapotranspiration (Jung et al., 2011) and baseflow (Beck et al. 2013) are already interpolated and extrapolated using machine learning. Moving

forwards, it is important to consider commensurability while applying machine learning in this context.

8) Consider comparing models against hydrologic signatures - indices that provide insight into the functional behavior of the system under study (Wagener et al., 2007; McMilan, 2020). The direct comparison of simulated and observed variables through statistical error metrics has at least two downsides. One, the above mentioned unresolved problem of commensurability, and two, the issue that such error metrics are rather uninformative in a diagnostic sense - simply knowing the size of an error does not tell the modeller how the model needs to be improved, only that it does (Yilmaz et al., 2009). One way to overcome these issues, is to derive hydrologically meaningful signatures from the original data, such as the signatures derived from transient groundwater levels by Heudorfer et al. (2019). For example, recharge ratio (defined as the ratio of groundwater recharge to precipitation) might be hydrologically more informative than recharge alone (Jasechko et al., 2014) or the water table ratio and groundwater response time (Cuthbert et al. 2019; Opie et al., 2020) which are spatially-distributed signatures of groundwater systems dynamics. Such signatures might be used to assess model consistency (Wagener & Gupta, 2005; Hrachowitz et al.2014) by looking at the similarity of patterns or spatial trends rather than the size of the aggregated error, thus reducing the commensurability problem.

9) Understand and quantify commensurability error issues better so that a fairer comparison can be made across scales using existing data. As described above, commensurability errors will depend on the number and locations of observation points, the variability structure of the variables being compared such as hydraulic head and the interpolation or aggregation scheme applied. While to some extent we may appreciate how each of these factors affect commensurability error in theory, in practice their combined effects are poorly understood and methods to quantify and reduce commensurability errors for groundwater model purposes remain largely undeveloped. As such, quantification of commensurability error in (large-scale) groundwater studies is regularly overlooked as a source of uncertainty because it cannot be satisfactorily evaluated (Tregoning et al., 2012). Currently, evaluation of simulated groundwater heads is plagued by, as yet, poorly quantified uncertainties stemming from commensurability errors and we therefore recommend future studies focus on developing solutions to this problem. An additional, subtle but important and unresolved commensurability issue can stem from conceptual models. Different hydrogeologists examining different scales, data or interpreting geology differently can produce quite different conceptual models of the same region (Troldborg et al. 2007).

We recommend evaluating models with a broader range of currently available data sources (with explicit consideration of data uncertainty and regional differences) while also simultaneously working to derive new data sets. Using data (such as baseflow, land subsidence, or the spatial distribution of perennial, intermittent, and ephemeral streams) that is more consistent with the scale modelled grid resolution will hopefully reduce the commensurability challenges. However, data distribution and commensurability issues will likely still be present, which underscores the importance of the two following strategies.

3.2. Model-based model evaluation

Model-based model evaluation, which includes model intercomparison projects (MIP) and model sensitivity and uncertainty analysis, can be done with or without explicitly using observations. We describe both inter-model and inter-scale comparisons which could be leveraged to maximize the strengths of each of these approaches.

The original MIP concept offers a framework to consistently evaluate and compare models, and associated model input, structural, and parameter uncertainty under different objectives (e.g., climate change, model performance, human impacts and developments). Early model intercomparisons of groundwater models focused on nuclear waste disposal (SKI, 1984). Since the Project for the Intercomparison of Land-Surface Parameterization Schemes (PILPS; Sellers et al., 1993), the first large-scale MIP, the land surface modeling community has used MIPs to deepen understanding of land physical processes and to improve their numerical implementations at various scales from regional (e.g., Rhône-aggregation project; Boone et al., 2004) to global (e.g., Global Soil Wetness Project; Dirmeyer, 2011). Two examples of recent model intercomparison efforts illustrate the general MIP objectives and practice. First, ISIMIP (Schewe et al., 2014; Warszawski et al., 2014) assessed water scarcity at different levels of global warming. Second, IH-MIP2 (Kollet et al., 2017) used both synthetic domains and an actual watershed to assess fully-integrated hydrologic models because these cannot be validated easily by comparison with analytical solutions and uncertainty remains in the attribution of hydrologic responses to model structural errors. Model comparisons have revealed differences, but it is often unclear whether these stem from differences in the model structures, differences in how the parameters were estimated, or from other modelling choices (Duan et al., 2006). Attempts for modular modelling frameworks to enable comparisons (Wagener et al., 2001; Leavesley et al., 2002; Clark et al., 2008; Fenicia et al., 2011; Clark et al., 2015) or at least shared explicit modelling protocols and boundary conditions (Refsgaard et al., 2007; Ceola et al., 2015; Warszawski et al., 2014) have been proposed to reduce these problems.

Inter-scale model comparison - for example, comparing a global model to a regional-scale model - is a potentially useful approach which is emerging for surface hydrology models (Hattermann et al., 2017; Huang et al., 2017) and could be applied to large-scale models with groundwater representation. For example, declining heads and decreasing groundwater discharge have been compared between a calibrated regional-scale model (RRCA, 2003) and a global model (de Graaf et al., 2019). A challenge to inter-scale comparisons is that regional-scale models often have more spatially complex subsurface parameterizations because they have access to local data which can complicate model inter-comparison. Another approach which may be useful is running large-scale models over smaller (regional) domains at a higher spatial resolution (same as a regional-scale model) so that model structure influences the comparison less. In the future, various variables that are hard to directly observe at large scales but routinely simulated in regional-scale models such as baseflow or recharge could be used to evaluate large-scale models, although these flux estimates can contain large uncertainty. In this way, the output fluxes and intermediate spatial scale of regional models provide a bridge across the "river of incommensurability" between highly location-specific data such as well observations and the coarse resolution of large-scale models. In such an evaluation, the uncertainty of flux estimates and scale of aggregation are both important to consider. It is important to consider that regional-scale models are not necessarily or inherently more accurate than large-scale models since problems may arise from conceptualization, groundwater-surface water interactions, scaling issues, parameterization etc.

In order for a regional-scale model to provide a useful evaluation of a large-scale model, there are several important documentation and quality characteristics it should meet. At a bare minimum, the regional-scale model must be accessible and therefore meet basic replicability requirements including open and transparent input and output data and model code to allow large-scale modelers to run the model and interpret its output. Documentation through peer review, either through a scientific journal or agency such as the US Geological Survey, would be ideal. It is particularly important that the documentation discusses limitations, assumptions and uncertainties in the regional-scale model so that a large-scale modeler can be aware of potential weaknesses and guide their comparison accordingly. Second, the boundary conditions and/or parameters being evaluated need to be reasonably comparable between the regional- and large-scale models. For example, if the regional-scale model includes human impacts through groundwater pumping while the large-scale model does not, a comparison of baseflow between the two models may not be appropriate. Similarly, there needs to be consistency in the time period simulated between the two models. Finally, as with data-driven model evaluation, the purpose of the large-scale model needs to be consistent with the model-based evaluation; matching the hydraulic head of a regional-scale model, for instance, does not indicate that estimates of stream-aquifer exchange are valid. Ideally, we recommend developing a community database of regional-scale models that meet this criteria. It is important to note that Rossman & Zlotnik (2014) review 88 regional-scale models while a good example of such a repository is the California Groundwater Model Archive (https://ca.water.usgs.gov/sustainable-groundwater-management/california-groundwater-modeling.html).

In addition to evaluating whether models are similar in terms of their outputs, e.g. whether they simulate similar groundwater head dynamics, it is also relevant to understand whether the influence of controlling parameters are similar across models. This type of analysis provides insights into process controls as well as dominant uncertainties. Sensitivity analysis provides the mathematical tools to perform this type of model evaluation (Saltelli et al., 2008; Pianosi et al., 2016; Borgonovo et al., 2017). Recent applications of sensitivity analysis to understand modelled controls on groundwater related processes include the study by Reinecke et al.

(2019b) trying to understand parametric controls on groundwater heads and flows within a global groundwater model. Maples et al. (2020) demonstrated that parametric controls on groundwater recharge can be assessed for complex models, though over a smaller domain. As highlighted by both of these studies, more work is needed to understand how to best use sensitivity analysis methods to assess computationally expensive, spatially distributed and complex groundwater models across large domains (Hill et al., 2016). In the future, it would be useful to go beyond parameter uncertainty analysis (e.g. Reinecke et al. 2019b) to begin to look at all of the modelling decisions holistically such as the forcing data (Weiland et al., 2015) and digital elevation models (Hawker et al., 2018). Addressing this problem requires advancements in statistics (more efficient sensitivity analysis methods), computing (more effective model execution), and access to large-scale models codes (Hutton et al. 2016), but also better utilization of process understanding, for example to create process-based groups of parameters which reduces the complexity of the sensitivity analysis study (e.g. Hartmann et al., 2015; Reinecke et al., 2019b).

3.3 Expert-based model evaluation

A path much less traveled is expert-based model evaluation which would develop hypotheses of phenomena (and related behaviors, patterns or signatures) we expect to emerge from large-scale groundwater systems based on expert knowledge, intuition, or experience. In essence, this model evaluation approach flips the traditional scientific method around by using hypotheses to test the simulation of emergent processes from large-scale models, rather than using large-scale models to test our hypotheses about environmental phenomena. This might be an important path forward for regions where available data is very sparse or unreliable. The recent discussion by Fan et al. (2019) shows how hypotheses about large-scale behavior might be derived from expert knowledge gained through the study of smaller scale systems such as critical zone observatories. While there has been much effort to improve our ability to make hydrologic predictions in ungauged locations through the regionalization of hydrologic variables or of model parameters (Bloeschl et al., 2013), there has been much less effort to directly derive expectations of hydrologic behavior based on our perception of the systems under study.

Large-scale models could then be evaluated against such hypotheses, thus providing a general opportunity to advance how we connect hydrologic understanding with large-scale modeling - a strategy that could also potentially reduce epistemic uncertainty (Beven et al., 2019), and which may be especially useful for groundwater systems given the data limitations described above. Developing appropriate and effective hypotheses is crucial and should likely focus on large-scale controlling factors or relationships between controlling factors and output in different parts of the model domain; hypotheses that are too specific may only be able to be tested by certain model complexities or in certain regions. To illustrate the type of hypotheses we are suggesting, we list some examples of hypotheses drawn from current literature:

- water table depth and lateral flow strongly affect transpiration partitioning (Famiglietti and Wood, 1994; Salvucci and Entekhabi, 1995; Maxwell & Condon, 2016);

- the percentage of inter-basinal regional groundwater flow increases with aridity or decreases with frequency of perennial streams (Gleeson & Manning, 2008; Goderniaux et al, 2013; Schaller and Fan, 2008); or
- human water use systematically redistributes water resources at the continental scale via non-local atmospheric feedbacks (Al-Yaari et al., 2019; Keune et al., 2018).

Alternatively, it might be helpful to also include hypotheses that have been shown to be incorrect since models should also not show relationships that have been shown to not exist in nature. For example of a hypotheses that has recently been shown to be incorrect is that the baseflow fraction (baseflow volume/precipitation volume) follows the Budyko curve (Gnann et al. 2019) . As yet another alternative, hydrologic intuition could form the basis of model experiments, potentially including extreme model experiments (far from the natural conditions). For example, an experiment that artificially lowers the water table by decreasing precipitation (or recharge directly) could hypothesize the spatial variability across a domain regarding how 'the drainage flux will increase and evaporation flux will decrease as the water table is lowered'. These hypotheses are meant only for illustrative purposes and we hope future community debate will clarify the most appropriate and effective hypotheses. We believe that the debate around these hypotheses alone will lead to advance our understanding, or, at least highlight differences in opinion.

Formal approaches are available to gather the opinions of experts and to integrate them into a joint result, often called expert elicitation (Aspinall, 2010; Cooke, 1991; O'Hagan, 2019). Expert elicitation strategies have been used widely to describe the expected behavior of environmental or man-made systems for which we have insufficient data or knowledge to build models directly. Examples include aspects of future sea-level rise (Bamber and Aspinall, 2013), tipping points in the Earth system (Lenton et al., 2018), or the vulnerability of bridges to scour due to flooding (Lamb et al., 2017). In the groundwater community, expert opinion is already widely used to develop system conceptualizations and related model structures (Krueger et al., 2012; Rajabi et al., 2018; Refsgaard et al., 2007), or to define parameter priors (Ross et al., 2009; Doherty and Christensen, 2011; Brunner et al., 2012; Knowling and Werner, 2016; Rajabi and Ataie-Ashtiani, 2016). The term expert opinion may be preferable to the term expert knowledge because it emphasizes a preliminary state of knowledge (Krueger et al., 2012).

A critical benefit of expert elicitation is the opportunity to bring together researchers who have experienced very different groundwater systems around the world. It is infeasible to expect that a single person could have gained in-depth experience in modelling groundwater in semi-arid regions, in cold regions, in tropical regions etc. Being able to bring together different experts who have studied one or a few of these systems to form a group would certainly create a whole that is bigger than the sum of its parts. If captured, it would be a tremendous source of knowledge for the evaluation of large-scale groundwater models. Expert elicitation also has a number of challenges including: 1) formalizing this knowledge in such a way that it is still usable by third parties that did not attend the expert workshop itself; and 2) perceived or real differences in perspectives, priorities and backgrounds between regional-scale and large-scale modelers.

So, while expert opinion and judgment play a role in any scientific investigation (O'Hagan, 2019), including that of groundwater systems, we rarely use formal strategies to elicit this opinion. It is also less common to use expert opinion to develop hypotheses about the dynamic behavior of groundwater systems, rather than just priors on its physical characteristics. Yet, it is intuitive that information about system behavior can help in evaluating the plausibility of model outputs (and thus of the model itself). This is what we call expert-based evaluation herein. Expert elicitation is typically done in workshops with groups of a dozen or so experts (e.g. Lamb et al., 2018). Upscaling such expert elicitation in support of global modeling would require some web-based strategy and a formalized protocol to engage a sufficiently large number of people. Contributors could potentially be incentivized to contribute to the web platform by publishing a data paper with all contributors as co-authors and a secondary analysis paper with just the core team as coauthors. We recommend the community develop expert elicitation strategies to identify effective hypotheses that directly link to the relevant large-scale hydrologic processes of interest.

4. CONCLUSIONS: towards a holistic evaluation of groundwater representation in large-scale models

Ideally, all three strategies (observation-based, model-based, expert-based) should be pursued simultaneously because the strengths of one strategy might further improve others. For example, expert- or model-based evaluation may highlight and motivate the need for new observations in certain regions or at new resolutions. Or observation-based model evaluation could highlight and motivate further model development or lead to refined or additional hypotheses. We thus recommend the community significantly strengthens efforts to evaluate large-scale models using all three strategies. Implementing these three model evaluation strategies may require a significant effort from the scientific community, so we therefore conclude with two tangible community-level initiatives that would be excellent first steps that can be pursued simultaneously with efforts by individual research groups or collaborations of multiple research groups.

First, we need to develop a 'Groundwater Modeling Data Portal' that would both facilitate and accelerate the evaluation of groundwater representation in continental to global scale models (Bierkens, 2015). Existing initiatives such as IGRAC's Global Groundwater Monitoring Network (https://www.un-igrac.org/special-project/ggmn-global-groundwater-monitoring-network) and HydroFrame (www.hydroframe.org), are an important first step but were not designed to improve the evaluation of large-scale models and the synthesized data remains very heterogeneous - unfortunately, even groundwater level time series data often remains either hidden or inaccessible for various reasons. This open and well documented data portal should include:

a) observations for evaluation (Table 2) as well as derived signatures (Section 3.1);

b) regional-scale models that meet the standards described above and could facilitate inter-scale comparison (Section 3.2) and be a first step towards linking regional models (Section 2.2);

c) Schematizations, conceptual or perceptual models of large-scale models since these are the basis of computational models; and

d) Hypothesis and other results derived from expert elicitation (Section 3.3).

Meta-data documentation, data tagging, aggregation and services as well as consistent data structures using well-known formats (netCDF, .csv, .txt) will be critical to developing a useful, dynamic and evolving community resource. The data portal should be directly linked to harmonized input data such as forcings (climate, land and water use etc.) and parameters (topography, subsurface parameters etc.), model codes, and harmonized output data. Where possible, the portal should follow established protocols, such as the Dublin Core Standards for metadata (https://dublincore.org) and ISIMIP protocols for harmonizing data and modeling approach, and would ideally be linked to or contained within an existing disciplinary repository such as HydroShare (https://www.hydroshare.org/) to facilitate discovery, maintenance, and long-term support. Additionally, an emphasis on model objective, uncertainty and regional differences as highlighted (Section 3) will be important in developing the data portal. Like expert-elicitation, contribution to the data portal could be incentivized through co-authorship in data papers and by providing digital object identifiers (DOIs) to submitted data and models so that they are citable. By synthesizing and sharing groundwater observations, models, and hypotheses, this portal would be broadly useful to the hydrogeological community beyond just improving global model evaluation.

Second, we suggest ISIMIP, or a similar model intercomparison project, could be harnessed as a platform to improve the evaluation of groundwater representation in continental to global scale models. For example, in ISIMIP (Warszawski et al., 2014), modelling protocols have been developed with an international network of climate-impact modellers across different sectors (e.g. water, agriculture, energy, forestry, marine ecosystems) and spatial scales. Originally, ISIMIP started with multi-model comparison (model-based model evaluation), with a focus on understanding how model projections vary across different sectors and different climate change scenarios (ISIMIP Fast Track). However, more rigorous model evaluation came to attention more recently with ISIMIP2a, and various observation data, such as river discharge (Global Runoff Data Center), terrestrial water storage (GRACE), and water use (national statistics), have been used to evaluate historical model simulation (observation-based model evaluation). To better understand model differences and to quantify the associated uncertainty sources, ISIMIP2b includes evaluating scenarios (land use, groundwater use, human impacts, etc) and key assumptions (no explicit groundwater representation, groundwater availability for the future, water allocation between surface water and groundwater), highlighting that different types of hypothesis derived as part of the expert-based model evaluation could possibly be simulated as part of the ISIMIP process in the future. While there has been a significant amount of research and publications on MIPs including surface water availability, limited multi-model assessments for large-scale groundwater studies exist. Important aspects of MIPs in general could facilitate all three model evaluation strategies: community-building and cooperation with various scientific communities and research groups, and making the model input and output publicly available in a standardized format.

Large-scale hydrologic and land surface models increasingly represent groundwater, which we envision will lead to a better understanding of large-scale water systems and to more sustainable water resource use. We call on various scientific communities to join us in this effort to improve the evaluation of groundwater in continental to global models. As described by examples above, we have already started this journey and we hope this will lead to better outcomes especially for the goals of including groundwater in large-scale models that we started with above: improving our understanding of Earth system processes; and informing water decisions and policy. Along with the community currently directly involved in large-scale groundwater modeling, above we have made pointers to other communities who we hope will engage to accelerate model evaluation: 1) regional hydrogeologists, who would be useful especially in expert-based model evaluation (Section 3.3); 2) data scientists with expertise in machine learning, artificial intelligence etc. whose methods could be useful especially for observation- and model-based model evaluation (Sections 3.1 and 3.2); and 3) the multiple Earth Science communities that are currently working towards integrating groundwater into a diverse range of models so that improved evaluation approaches are built directly into model development. Together we can better understand what has always been beneath our feet, but often forgotten or neglected.

Competing interests: The authors declare that they have no conflict of interest.

Acknowledgements:

The commentary is based on a workshop at the University of Bristol and significant debate and discussion before and after. This community project was directly supported by a Benjamin Meaker Visiting Professorship at the Bristol University to TG and by funding from the Alexander von Humboldt Foundation to TW in the framework of the Alexander von Humboldt Professorship endowed by the German Federal Ministry of Education and Research. We thank many members of the community who contributed to the discussions, especially at the IGEM (Impact of Groundwater in Earth System Models) workshop in Taiwan.

Author Contributions: (using the CRediT taxonomy which offers standardized descriptions of author contributions) conceptualization and writing original draft: TG, TW and PD; writing -

review and editing:all co-authors. Authors are ordered by contribution for the first three coauthors (TG, TW and PD) and then ordered in reverse alphabetical order for all remaining coauthors.

Code and data availability: This Perspective paper does not present any computational results. There is therefore no code or data associated with this paper.

Table 1. A possible model classification based on three model classes and various model characteristics; see link to google doc to view easier (google doc will be migrated to a community github page if article accepted)

Table 1. Model classification for large-scale models representing groundwater (1)

groundwater flow	No GW flow	lateral groundwater flow to a river within a cell						2D lateral groundwater flow between all cells				3D groundwater flow	
groundwater-surface coupling (2)		one-way		two-way				one-way		two-way			
surface-atmosphere coupling	yes							no					
example model (3)	JULES	ORCHIDEE	LM3	VIC-ground	CLM5	TOPLATS	Catchment	WaterGAP2-G3M	LEAF hydro	PCRGLOB-WB - MODFLOW	ISBA-TRIP	HydroGeoSphere	ParFlow
groundwater recharge (diffuse)	Free-drainage	Recharge = P-R-ET	Recharge = P-R-ET	Recharge depends on WT head and capillary fluxes	Recharge depends on WT head and capillary fluxes	Recharge depends on WT head and capillary fluxes	Recharge depends on WT head and capillary fluxes	currently uncoupled	recharge derived from WGM	Recharge depends on WT head and capillary fluxes	Recharge depends on WT head and capillary fluxes	directly represented	directly represented
focused recharge (4)	not represented	optional (via enhanced infiltration in ponds)	not represented	not represented	not represented	not represented	not represented	represented after coupling	not represented	represented from lakes and perenial rivers?	not represented	not represented	not represented
surface water boundary condition or coupling	not represented	not represented	not represented	not represented	not represented	not represented	not represented	currently uncoupled with boundary condition using conductance	no head-based interactions with surface water	one-way coupling with three boundary conditions including drainage from linear reservoir	directly represented	directly represented	directly represented
variably saturated or partially saturated (5)	1D Richards' in soil layers	1D Richards' in soil layers	1D Richards' in soil layers	1D Richards' in soil layers	1D Richards' in soil layers	1D Richards' in soil layers	Lumped 3D Richards	partially saturated	partially saturated	vertical fluxes in soils depend ing on soil saturation and GW level	1D Richards' in soil layers	variably saturated using 3D Richard's equation	variably saturated using 3D Richard's equation
water table and hydraulic head	Optional WT diagnostic based on TOPMODEL	not represented	represented, parameterized	directly represented	First layer from bedrock where soil moisture < 0.8	represented following TOPMODEL	represented following TOPMODEL	directly represented	directly represented	directly represented	directly represented	directly represented	directly represented
groundwater storage	not represented	represented as linear reservoir	represented	represented	represented	represented	represented	directly represented	represented	directly represented	directly represented	directly represented	directly represented
lateral flow	not represented	represented	represented through lateral flow divergence	parametrized following Francini and Pacciani (2001)	parameterised, calibration parameter related to baseflow	represented following TOPMODEL	represented following TOPMODEL	directly represented but not along flowlines	directly represented	directly represented	directly represented	directly represented	directly represented
groundwater bottom boundary condition	gravity drainage from soil	function of reservoir	no flux	no flux	no flux	no flux	no flux	no flux	no flux	no flux	no flux	no flux	no flux
groundwater use	not represented	not represented	not represented	not represented	not represented	not represented	not represented	to be included in future	not represented	represented	not represented	not represented	not represented
preferential flow	not represented	not represented	not represented	not represented	not represented	not represented	not represented	not represented	not represented	not represented	not represented	not represented	not represented
groundwater temperature	not represented	not represented	not represented	not represented	not represented	not represented	not represented	not represented	not represented	not represented	not represented	not represented	not represented
groundwater quality	not represented	not represented	not represented	not represented	not represented	not represented	not represented	not represented	not represented	not represented	not represented	not represented	not represented
groundwater density	not represented	not represented	not represented	not represented	not represented	not represented	not represented	not represented	not represented	not represented	not represented	not represented	not represented
confined conditions	not represented	not represented	not represented	not represented	not represented	not represented	not represented	not represented	not represented	represented	not represented	not represented	not represented
coupling with ocean (and ocean models)	no	no	no	no	no	no	no	no	ocean boundary condition	ocean boundary condition	???	ocean boundary condition	possible
isotope-enabled	no	no	no	no	no	no	no	no	no	no	no	no	no
included in current assimilation schemes	yes	???	no	no	yes	???	no	no	no	no	no	no	no
paleo groundwater	not represented	not represented	not represented	not represented	not represented	not represented	not represented	not represented	not represented	not represented	not represented	not represented	not represented
Reference	Best et al. (2011)	Guimberteau et al. (	Milly et al (2014)	Liang et al. (2003	Andre et al. (2018	Famiglietti & Wood (	Koster et al. (2000)	Reinecke et al. (201	Fan et al. (2013	de Graaf et al. (2017	Vergnes et al. (2014)	Brunner and Simmo	Maxwell et al (2017

Notes:

(1) Only the most RECENT version of models with published results at continental to global scales are included. Analytical solutions (including the water table ratio or groundwater response times) are not described here.
(2) one-way coupling means that SM -> recharge -> GW -> stream flow, but no reverse influence; in this case, the GW model is dependent on surface simulations to provide recharge. two-way coupling means there is a fully coupling of surf
(3) Other models exist with similar features
(4) Focused recharge refers to any recharge that occurs beneath water bodies such as streams or lakes; whereas preferential flow to mean recharge that bypasses the soil matrix during diffuse recharge through fractures or other macropores
(5) Variably saturated means that the saturation, and related constitutive relations can vary continiously, while partially saturated means that saturation can only discretely vary between fully saturated and unsaturated.

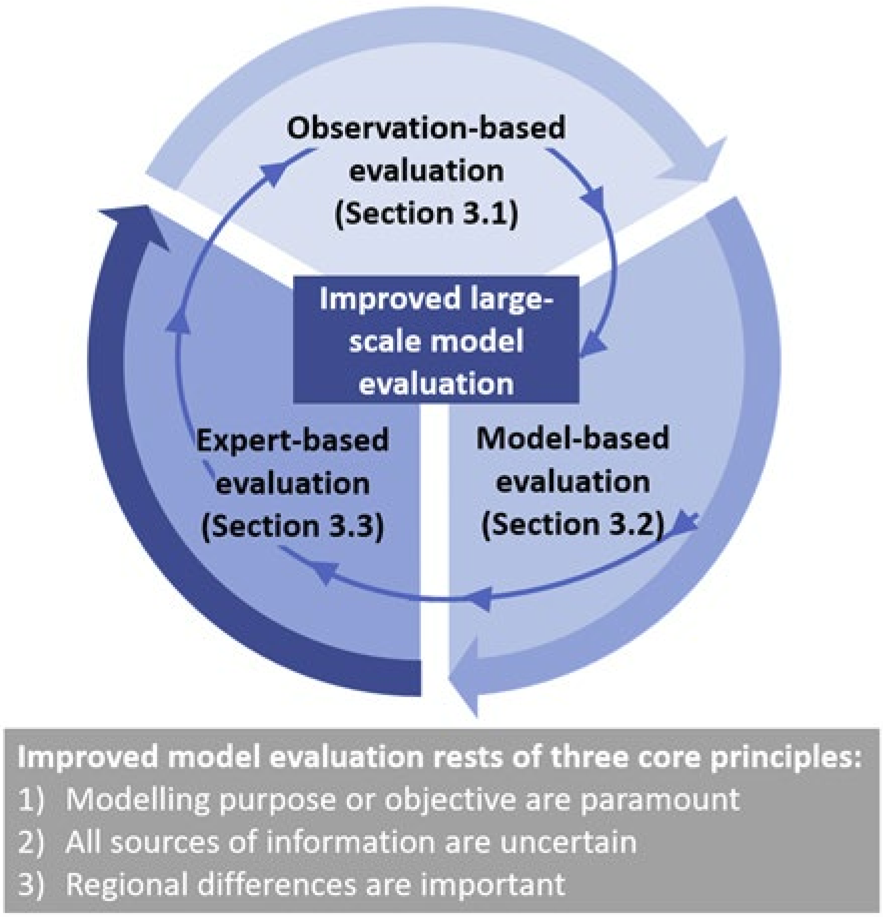
Table 2. Available observations for evaluating the groundwater component of large-scale models

Data type	Strengths	Limitations	Data availability and spatial resolution
Available observations already used to evaluate large-scale models			
Hydraulic heads or water table depth (averages or single times)	Direct observation of groundwater levels and storage	observations biased towards North America and Europe; non-commensurable with large-scale models; mixture of observation times	IGRAC Global Groundwater Monitoring Network; USGS; Fan et al. (2013) Point measurements at existing wells
Hydraulic heads or water table depth (transient)	Direct observation of changing groundwater levels and storage	As above	time-series available in a few regions, especially through USGS and European Groundwater Drought Initiative Point measurements at existing wells
Total water storage anomalies (GRACE)	Globally available and regionally integrated signal of water storage trends and anomalies	Groundwater changes are uncertain model remainder; very coarse spatial resolution and limited period	Various mascons gridded with resolution of ~100,000 km^2 which are then processed as groundwater storage change; Scanlon et al. (2016)
Storage change (regional aquifers)	Regionally integrated response of aquifer (independent estimates derived by various methods)	Bias towards North America and Europe	Konikow (2011); Döll et al. (2014a) Regional aquifers (10,000s to 100,000s km^2)
Recharge	Direct inflow of groundwater system	Challenging to measure and upscale	Döll and Fiedler (2008); Hartmann et al. (2017); Mohan et al. (2018); Moeck et al. (2020)

			Point to small basin
Abstractions	Crucial for groundwater depletion and sustainability studies	National scale data highly variable in quality; downscaling uncertain	de Graaf et al. (2014); Döll et al. (2014a) National-scale data down-scaled to grid
Streamflow or spring flow observations	Widely available at various scales; low flows can be related to groundwater	Challenging to quantify the flows between groundwater and surface water from streamflow	Global Runoff Data Centre (GRDC) or other data sources; large to small basin; Olarinoye et al. (2020) point measurements of spring flow
Evapotranspiration	Widely available; related to groundwater recharge or discharge (for shallow water tables)	Not a direct groundwater observations	Various datasets; e.g. Miralles et al. (2016); gridded
Available observations not being used to evaluate large-scale models			
Baseflow index (BFI) or (non-)linear baseflow recession behavior	Possible integrator of groundwater contribution to streamflow over a basin	BFI and k values vary with method; baseflow may be dominated by upstream surface water storage rather than groundwater inflow; can not identify losing river conditions	Beck et al. (2013) Point observations extrapolated by machine learning

Perennial stream map	Ephemeral streams are losing streams, whereas perennial streams could be gaining (or impacted by upstream surface water storage)	Mapping perennial streams requires arbitrary streamflow and duration cutoffs; not all perennial stream reaches are groundwater-influenced; does not provide information about magnitude of inflows/outflows.	Schneider et al. (2017); Cuthbert et al. (2019); Spatially continuous along stream networks
Gaining or losing stream reaches	Multiple techniques for measurement (interpolated head measurements, streamflow data, water chemistry). Constrains direction of fluxes at groundwater system boundaries	Relevant processes occur at sub-grid-cell resolution.	Not globally available but see Bresciani et al. (2018) for a regional example; Spatially continuous along stream networks
Springs and groundwater-dependent surface water bodies	Constrains direction of fluxes at groundwater system boundaries	Relevant processes occur at sub-grid-cell resolution.	Springs available for various regions but not globally; Springer, & Stevens (2009) Point measurements at water feature locations
Tracers (heat, isotopes or other geochemical)	Provides information about temporal aspects of groundwater systems (e.g. residence time)	No large-scale models simulate transport processes (Table S1)	Isotopic data compiled but no global data for heat or other chemistry; Gleeson et al. (2016); Jasechko et al. (2017) Point measurements at existing wells or surface water features
Surface elevation data (leveling, GPS, radar/lidar) an in particular land subsidence observations	Provides information about changes in surface elevation that are related to groundwater head variations or groundwater head decline	Provides indirect information and needs a geomechanical model to translate to head. Introduces additional uncertainty of geomechnical properties.	Leveling data, GPS data and lidar observations mostly limited to areas of active subsidence; Minderhoud et al. (2019,2020). Global data on elevation change are available from the Sentinel 1 mission.

Figure 1: Improved large-scale model evaluation rests on three pillars: observation-, model-, and expert-based model evaluation. We argue that each pillar is an essential strategy so that all three should be simultaneously pursued by the scientific community. The three pillars of model evaluation all rest on three core principles related to 1) model objectives, 2) uncertainty and 3) regional differences.



References

Addor, N., & Melsen, L. A. (2018). Legacy, Rather Than Adequacy, Drives the Selection of Hydrological Models. *Water Resources Research*, *0*(0). https://doi.org/10.1029/2018WR022958

Al-Yaari, A., Ducharne, A., Cheruy, F., Crow, W.T. & Wigneron, J.P. (2019). Satellite-based soil moisture provides missing link between summertime precipitation and surface temperature biases in CMIP5 simulations over conterminous United States. *Scientific Reports*, 9, article number 1657, doi:10.1038/s41598-018-38309-5

Anderson, M. P., Woessner, W. W. & Hunt, R. (2015a). *Applied groundwater modeling- 2nd Edition*. San Diego: Academic Press.

Anderson, R. G., Min-Hui Lo, Swenson, S., Famiglietti, J. S., Tang, Q., Skaggs, T. H., Lin, Y.-H., and Wu, R.-J. (2015b), Using satellite-based estimates of evapotranspiration and groundwater changes to determine anthropogenic water fluxes in land surface models, Geosci. Model Dev., 8, 3021-3031, doi:10.5194/gmd-8-3021-2015.Alley, W.M. and LF Konikow (2015) Bringing GRACE down to earth. Groundwater 53 (6): 826–829

Anyah, R. O., Weaver, C. P., Miguez-Macho, G., Fan, Y., & Robock, A. (2008). Incorporating water table dynamics in climate modeling: 3. Simulated groundwater influence on coupled land-atmosphere variability. *J. Geophys. Res.*, *113*. Retrieved from http://dx.doi.org/10.1029/2007JD009087

Archfield, S. A., Clark, M., Arheimer, B., Hay, L. E., McMillan, H., Kiang, J. E., et al. (2015). Accelerating advances in continental domain hydrologic modeling. *Water Resources Research*, *51*(12), 10078–10091. https://doi.org/10.1002/2015WR017498

Aspinall, W. (2010). A route to more tractable expert advice. *Nature*, *463*, 294–295. https://doi.org/10.1038/463294a

ASTM (2016), Standard Guide for Conducting a Sensitivity Analysis for a Groundwater Flow Model Application, ASTM International D5611-94, West Conshohocken, PA, 2016, www.astm.org

Bamber, J.L. and Aspinall, W.P. (2013). An expert judgement assessment of future sea level rise from the ice sheets. Nature Climate Change. 3(4), 424-427.

Barnett, B., Townley, L.R., Post, V.E.A., Evans, R.E., Hunt, R.J., Peeters, L., Richardson, S., Werner, A.D., Knapton, A., Boronkay, A. (2012). Australian groundwater modelling guidelines, National Water Commission, Canberra, 203 pages

Barthel, R. (2014). HESS Opinions "Integration of groundwater and surface water research: an interdisciplinary problem?" *Hydrology and Earth System Sciences*, *18*(7), 2615–2628.

Beck, H. et al (2013). Global patterns in base flow index and recession based on streamflow observations from 3394 catchments. Water Resources Research.

Befus, K., Jasechko, S., Luijendijk, E., Gleeson, T., Cardenas, M.B. (2017) The rapid yet uneven turnover of Earth's groundwater. (2017) Geophysical Research Letters 11: 5511-5520 doi: 10.1002/2017GL073322

Best, M. J., Pryor, M., Clark, D. B., Rooney, G. G., Essery, R. L. H., Ménard, C. B., Edwards, J. M., Hendry, M. A., Porson, A., Gedney, N., Mercado, L. M., Sitch, S., Blyth, E., Boucher, O., Cox, P. M., Grimmond, C. S. B., & Harding, R. J. (2011). The Joint UK Land Environment Simulator (JULES), model description – Part 1: Energy and water fluxes, *Geosci. Model Dev*., 4, 677-699. https://doi.org/10.5194/gmd-4-677-2011

Beven, K. (2000). Uniqueness of place and process representations in hydrological modelling. Hydrology and Earth System Sciences, 4(2), 203–213.

Beven, K. (2005). On the concept of model structural error. Water Science & Technology, 52(6), 167–175.

Beven, K. (2016). Facets of uncertainty: epistemic uncertainty, nonstationarity, likelihood, hypothesis testing, and communication. Hydrological Sciences Journal, 61(9), 1652-1665, DOI: 10.1080/02626667.2015.1031761

Beven, K. (2019) How to make advances in hydrological modelling. In: Hydrology Research. 50, 6, p. 1481-1494. 14 p.

Beven, K. (2020). Deep learning, hydrological processes and the uniqueness of place. Hydrological Processes, 34(16), 3608–3613. https://doi.org/10.1002/hyp.13805

Beven, K. J., and H. L. Cloke (2012), Comment on ‘‘Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth’s terrestrial water’’ by Eric F. Wood et al., Water Resour.Res., 48, W01801, doi:10.1029/2011WR010982.

Beven, K.J., Aspinall, W.P., Bates, P.D., Borgomeo, E., Goda, K., Hall, J.W., Page, T., Phillips, J.C., Simpson, M., Smith, P.J., Wagener, T. and Watson, M. 2018. Epistemic uncertainties and natural hazard risk assessment – Part 2: What should constitute good practice? Natural Hazards and Earth System Sciences, 18, 10.5194/nhess-18-1-2018

Bierkens, M. F. P. (2015). Global hydrology 2015: State, trends, and directions. *Water Resources Research*, *51*(7), 4923–4947. https://doi.org/10.1002/2015WR017173

Bierkens, M. F.P. & Wada, Y. (2019). Non-renewable groundwater use and groundwater depletion: A review. Environmental Research Letters, 14(6), 063002

Boone, A. A., Habets, F., Noilhan, J., Clark, D., Dirmeyer, P., Fox, S., Gusev, Y., Haddeland, I., Koster, R., Lohmann, D., Mahanama, S., Mitchell, K., Nasonova, O., Niu, G. Y., Pitman, A., Polcher, J., Shmakin, A. B., Tanaka, K., Van Den Hurk, B., Vérant, S., Verseghy, D., Viterbo, P. and Yang, Z. L.: The Rhône-aggregation land surface scheme intercomparison project: An overview, J. Clim., 17(1), 187–208, doi:10.1175/1520-0442(2004)017<0187:TRLSSI>2.0.CO;2, 2004.

Borgonovo, E. Lu, X. Plischke, E. Rakovec, O. and Hill, M. C. (2017). Making the most out of a hydrological model data set: Sensitivity analyses to open the model black-box. Water Resources Research. DOI:10.1002/2017WR020767

Bresciani, E., P. Goderniaux, and O. Batelaan (2016), Hydrogeological controls of water table-land surface interactions, Geophysical Research Letters, 43, 9653-9661.

Bresciani, E., Cranswick, R. H., Banks, E. W., Batlle-Aguilar, J., et al. (2018). Using hydraulic head, chloride and electrical conductivity data to distinguish between mountain-front and mountain-block recharge to basin aquifers. *Hydrology and Earth System Sciences*, 22(2), 1629–1648.

Brunner, P., J. Doherty, and C. T. Simmons (2012), Uncertainty assessment and implications for data acquisition in support of integrated hydrologic models, *Water Resources Research*, *48*.

Burgess, W. G., Shamsudduha, M., Taylor, R. G., Zahid, A., Ahmed, K. M., Mukherjee, A., et al. (2017). Terrestrial water load and groundwater fluctuation in the Bengal Basin. *Scientific Reports*, *7*(1), 3872.

Caceres, D., Marzeion, B., Malles, J.H., Gutknecht, B., Müller Schmied, H., Döll, P. (2020): Assessing global water mass transfers from continents to oceans over the period 1948–2016. Hydrol. Earth Syst. Sci. Discuss. doi:10.5194/hess-2019-664

Ceola, S., Arheimer, B., Baratti, E., Blöschl, G., Capell, R., Castellarin, A., et al. (2015). Virtual laboratories: new opportunities for collaborative water science. *Hydrology and Earth System Sciences*, *19*(4), 2101–2117.

Clark, M. P., A. G. Slater, D. E. Rupp, R. A. Woods, J. A. Vrugt, H. V. Gupta, T. Wagener, and L. E. Hay (2008) Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models, Water Resour. Res., 44, W00B02, doi:10.1029/2007WR006735.

Clark, M. P., et al. (2015), A unified approach for process-based hydrologic modeling: 1. Modeling concept,Water Resources Research,51, 2498–2514,doi:10.1002/2015WR017198

Condon, L. E., & Maxwell, R. M. (2019). Simulating the sensitivity of evapotranspiration and streamflow to large-scale groundwater depletion. *Science Advances*, *5*(6), eaav4574. https://doi.org/10.1126/sciadv.aav4574

Condon, LE et al Evapotranspiration depletes groundwater under warming over the contiguous United States Nature Comm, 2020, https://doi.org/10.1038/s41467-020-14688-0

Condon, L. E., Markovich, K. H., Kelleher, C. A., McDonnell, J. J., Ferguson, G., & McIntosh, J. C. (2020). Where Is the Bottom of a Watershed? Water Resources Research, 56(3). https://doi.org/10.1029/2019wr026010

Condon, L.E., Stefan Kollet, Marc F.P. Bierkens, Reed M. Maxwell, Mary C. Hill, Anne Verhoef, Anne F. Van Loon, Graham E. Fogg, Mauro Sulis , Harrie-Jan Hendricks Fransen ; Corinna Abesser. Global groundwater modeling and monitoring?: Opportunities and challenges (in review at WRR)

Cooke, R. (1991). *Experts in uncertainty: opinion and subjective probability in science*. Oxford University Press on Demand.

Cuthbert, M. O., Gleeson, T., Moosdorf, N., Befus, K. M., Schneider, A., Hartmann, J., & Lehner, B. (2019). Global patterns and dynamics of climate–groundwater interactions. *Nature Climate Change*, 9, 137–141 https://doi.org/10.1038/s41558-018-0386-4

Cuthbert, M. O., et al. (2019) Observed controls on resilience of groundwater to climate variability in sub-Saharan Africa. Nature 572: 230–234

Dalin, C., Wada, Y., Kastner, T., & Puma, M. J. (2017). Groundwater depletion embedded in international food trade. *Nature*, *543*(7647), 700–704. https://doi.org/10.1038/nature21403

DeAngelis, A., Dominguez, F., Fan, Y., Robock, A., Kustu, M. D., & Robinson, D. (2010). Evidence of enhanced precipitation due to irrigation over the Great Plains of the United States. Journal of Geophysical Research: Atmospheres, 115(D15).

Dirmeyer, P. A.: A History and Review of the Global Soil Wetness Project (GSWP), J. Hydrometeorol., 12(5), 110404091221083, doi:10.1175/jhm-d-10-05010, 2011

Doherty, J., and S. Christensen (2011), Use of paired simple and complex models to reduce predictive bias and quantify uncertainty, *Water Resources Research*, *47*(12),

Döll, P., Fiedler, K. (2008): Global-scale modeling of groundwater recharge. Hydrol. Earth Syst. Sci., 12, 863-885, doi: 10.5194/hess-12-863-2008

Döll, P., Douville, H., Güntner, A., Müller Schmied, H., Wada, Y. (2016): Modelling freshwater resources at the global scale: Challenges and prospects. Surveys in Geophysics, 37(2), 195-221. doi: 10.1007/s10712-015-9343-1

Döll, P., Müller Schmied, H., Schuh, C., Portmann, F. T., & Eicker, A. (2014a). Global-scale assessment of groundwater depletion and related groundwater abstractions: Combining hydrological modeling with information from well observations and GRACE satellites. *Water Resources Research*, *50*(7), 5698–5720. https://doi.org/10.1002/2014WR015595

Döll, P., Fritsche, M., Eicker, A., Müller Schmied, H. (2014b): Seasonal water storage variations as impacted by water abstractions: Comparing the output of a global hydrological model with GRACE and GPS observations. Surveys in Geophysics, 35(6), 1311-1331, doi: 10.1007/s10712-014-9282-2.

Döll, P., Hoffmann-Dobrev, H., Portmann, F.T., Siebert, S., Eicker, A., Rodell, M., Strassberg, G., Scanlon, B. (2012): Impact of water withdrawals from groundwater and surface water on continental water storage variations. J. Geodyn. 59-60, 143-156, doi:10.1016/j.jog.2011.05.001.

Duan Q., Schaake, J., Andreassian, V., Franks, S., Gupta, H.V., Gusev, Y.M., Habets, F., Hall, A., Hay, L., Hogue, T.S., Huang, M., Leavesley, G., Liang, X., Nasonova, O.N., Noilhan, J., Oudin, L., Sorooshian, S., Wagener, T. and Wood, E.F. (2006). Model Parameter Estimation Experiment (MOPEX): Overview and Summary of the Second and Third Workshop Results. *Journal of Hydrology*, 320(1-2), 3-17.

Enemark, T., Peeters, L. J. M., Mallants, D., & Batelaan, O. (2019). Hydrogeological conceptual model building and testing: A review. *Journal of Hydrology*, *569*, 310–329. https://doi.org/10.1016/j.jhydrol.2018.12.007

Erban L E, Gorelick S M and Zebker H A 2014 Groundwater extraction, land subsidence, and sea-level rise in the Mekong Delta, Vietnam Environ. Res. Lett. 9 084010

Famiglietti, J. S., & E. F. Wood (1994). Multiscale modeling of spatially variable water and energy balance processes, Water Resour. Res., 30(11), 3061–3078, https://doi.org/10.1029/94WR01498

Fan, Y. et al., (2019) Hillslope hydrology in global change research and Earth System modeling. *Water Resources Research*, doi.org/10.1029/2018WR023903

Fan, Y. (2015). Groundwater in the Earth's critical zone: Relevance to large-scale patterns and processes. *Water Resources Research*, *51*(5), 3052–3069. https://doi.org/10.1002/2015WR017037

Fan, Y., & Miguez-Macho, G. (2011). A simple hydrologic framework for simulating wetlands in climate and earth system models. *Climate Dynamics*, *37*(1–2), 253–278.

Fan, Y., Li, H., & Miguez-Macho, G. (2013). Global patterns of groundwater table depth. Science, 339(6122), 940–943.

Fenicia, F., D. Kavetski, and H. H. G. Savenije (2011), Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development, Water Resources Research, 47(11), W11510, 10.1029/2010wr010174.

Forrester, M.M. and Maxwell, R.M. Impact of lateral groundwater flow and subsurface lower boundary conditions on atmospheric boundary layer development over complex terrain. Journal of Hydrometeorology, doi:10.1175/JHM-D-19-0029.1, 2020.

Forrester, M.M., Maxwell, R.M., Bearup, L.A., and Gochis, D.J. Forest Disturbance Feedbacks from Bedrock to Atmosphere Using Coupled Hydro-Meteorological Simulations Over the Rocky Mountain Headwaters. Journal of Geophysical Research-Atmospheres, 123:9026-9046, doi:10.1029/2018JD028380 2018.

Freeze, R. A., & Witherspoon, P. A. (1966). Theoretical analysis of regional groundwater flow, 1. Analytical and numerical solutions to a mathematical model. *Water Resources Research*, *2*, 641–656.

Foster, S., Chilton, J., Nijsten, G.-J., & Richts, A. (2013). Groundwater — a global focus on the 'local resource.' *Current Opinion in Environmental Sustainability*, *5*(6), 685–695. doi.org/10.1016/j.cosust.2013.10.010

Garven, G. (1995). Continental-scale groundwater flow and geologic processes. *Annual Review of Earth and Planetary Sciences*, *23*, 89–117.

Gascoin, S., Ducharne, A., Ribstein, P., Carli, M., Habets, F. (2009). Adaptation of a catchment-based land surface model to the hydrogeological setting of the Somme River basin (France). *Journal of Hydrology*, 368(1-4), 105-116. https://doi.org/10.1016/j.jhydrol.2009.01.039

Genereux, D. (1998). Quantifying uncertainty in tracer-based hydrograph separations. *Water Resources Research*, *34*(4), 915–919.

Gilbert, J.M., Maxwell, R.M. and Gochis, D.J. Effects of water table configuration on the planetary boundary layer over the San Joaquin River watershed, California. Journal of Hydrometeorology, 18:1471-1488, doi:10.1175/JHM-D-16-0134.1, 2017.

Gleeson, T. et al. (2020) HESS Opinions: Improving the evaluation of groundwater representation in continental to global scale models. https://hess.copernicus.org/preprints/hess-2020-378/

Gleeson, T., & Manning, A. H. (2008). Regional groundwater flow in mountainous terrain: Three-dimensional simulations of topographic and hydrogeologic controls. *Water Resources Research*, *44*. Retrieved from http://dx.doi.org/10.1029/2008WR006848

Gleeson, T., Befus, K. M., Jasechko, S., Luijendijk, E., & Cardenas, M. B. (2016). The global volume and distribution of modern groundwater. *Nature Geosci*, *9*(2), 161–167.

de Graaf, I. E. M., van Beek, L. P. H., Wada, Y., & Bierkens, M. F. P. (2014). Dynamic attribution of global water demand to surface water and groundwater resources: Effects of abstractions and return flows on river discharges. *Advances in Water Resources*, *64*(0), 21–33. https://doi.org/10.1016/j.advwatres.2013.12.002

de Graaf, I. E. M., Sutanudjaja, E. H., Van Beek, L. P. H., & Bierkens, M. F. P. (2015). A high-resolution global-scale groundwater model. Hydrology and Earth System Sciences, 19(2), 823–837.

de Graaf, I. E. M., van Beek, L. P. H., Gleeson, T., Moosdorf, N., Schmitz, O., Sutanudjaja, E. H., & Bierkens, M. F. P. (2017). A global-scale two-layer transient groundwater model: Development and application to groundwater depletion. Advances in Water Resources, 102, 53–67. https://doi.org/10.1016/j.advwatres.2017.01.011

de Graaf, I. E. M., Gleeson, T., Beek, L. P. H. (Rens) van, Sutanudjaja, E. H., & Bierkens, M. F. P. (2019). Environmental flow limits to global groundwater pumping. Nature, 574(7776), 90–94. https://doi.org/10.1038/s41586-019-1594-4

Gnann, S. J., Woods, R. A., & Howden, N. J. (2019). Is there a baseflow Budyko curve? *Water Resources Research*, *55*(4), 2838–2855.

Goderniaux, P., P. Davy, E. Bresciani, J.-R. de Dreuzy, and T. Le Borgne (2013), Partitioning a regional groundwater flow system into shallow local and deep regional flow compartments, Water Resources Research, 49(4), 2274-2286.

Gosling, S. N., Zaherpour, J., Mount, N. J., Hattermann, F. F., Dankers, R., Arheimer, B., et al. (2017). A comparison of changes in river runoff from multiple global and catchment-scale hydrological models under global warming scenarios of 1 °C, 2 °C and 3 °C. *Climatic Change*, *141*(3), 577–595. https://doi.org/10.1007/s10584-016-1773-3

Guimberteau, M., Ducharne, A., Ciais, P., Boisier, J. P., Peng, S., De Weirdt, M., & Verbeeck, H. (2014). Testing conceptual and physically based soil hydrology schemes against observations for the Amazon Basin, Geosci. Model Dev., 7, 1115-1136. https://doi.org/10.5194/gmd-7-1115-2014

Habets, F., Boé, J., Déqué, M., Ducharne, A., Gascoin, S., Hachour, A., Martin, E., Pagé, C., Sauquet, E., Terray, L., Thiéry, D., Oudin, L. & Viennot, P. (2013). Impact of climate change on surface water and ground water of two basins in Northern France: analysis of the uncertainties associated with climate and hydrological models, emission scenarios and downscaling methods. *Climatic Change*, 121, 771-785. https://doi.org/10.1007/s10584-013-0934-x

Hartmann, A., Gleeson, T., Rosolem, R., Pianosi, F., Wada, Y., & Wagener, T. (2015). A large-scale simulation model to assess karstic groundwater recharge over Europe and the Mediterranean. *Geosci. Model Dev.*, *8*(6), 1729–1746. https://doi.org/10.5194/gmd-8-1729-2015

Hartmann, Andreas, Gleeson, T., Wada, Y., & Wagener, T. (2017). Enhanced groundwater recharge rates and altered recharge sensitivity to climate variability through subsurface heterogeneity. *Proceedings of the National Academy of Sciences*, *114*(11), 2842–2847. https://doi.org/10.1073/pnas.1614941114

Hattermann, F. F., Krysanova, V., Gosling, S. N., Dankers, R., Daggupati, P., Donnelly, C., et al. (2017). Cross-scale intercomparison of climate change impacts simulated by regional and global hydrological models in eleven large river basins. *Climatic Change*, *141*(3), 561–576. https://doi.org/10.1007/s10584-016-1829-4

Hay, L., Norton, P., Viger, R., Markstrom, S., Regan, R. S., & Vanderhoof, M. (2018). Modelling surface-water depression storage in a Prairie Pothole Region. Hydrological Processes, 32(4), 462–479. https://doi.org/10.1002/hyp.11416

Henderson-Sellers, A., Z. L. Yang, and R. E. Dickinson: The Project for Intercomparison of Land-Surface Schemes (PILPS). Bull. Amer. Meteor. Soc., 74, 1335–1349, 1993

Herbert, C., & Döll, P. (2019). Global assessment of current and future groundwater stress with a focus on transboundary aquifers. Water Resources Research, 55, 4760–4784.https://doi.org/10.1029/2018WR023321

Heudorfer, B., Haaf, E., Stahl, K., & Barthel, R. (2019). Index-based characterization and quantification of groundwater dynamics. Water Resources Research, 55, 5575–5592.https://doi.org/10.1029/2018WR024418

Hill, M. C. (2006). The practical use of simplicity in developing ground water models. *Ground Water*, *44*(6), 775–781. https://doi.org/10.1111/j.1745-6584.2006.00227.x

Hill, M. C., & Tiedeman, C. R. (2007). *Effective groundwater model calibration*. Wiley.

Hill, M. C., Kavetski, D. Clark, M. Ye, M. Arabi, M. Lu, D. Foglia, L. & Mehl, S. (2016). Practical use of computationally frugal model analysis methods. Groundwater. DOI:10.1111/gwat.12330

Hiscock, K. M., & Bense, V. F. (2014). *Hydrogeology—principles and practice* (2nd edition). Blackwell.

Huang, S., Kumar, R., Flörke, M., Yang, T., Hundecha, Y., Kraft, P., et al. (2017). Evaluation of an ensemble of regional hydrological models in 12 large-scale river basins worldwide. *Climatic Change*, *141*(3), 381–397. https://doi.org/10.1007/s10584-016-1841-8

Hrachowitz, M., Fovet, O., Ruiz, L., Euser, T., Gharari, S., Nijzink, R., Freer, J., Savenije, H.H.G. and Gascuel-Odoux, C. (2014). Process Consistency in Models: the Importance of System Signatures, Expert Knowledge and Process Complexity. Water Resources Research 50:7445-7469.

Hunt, R. J., Walker, J. F., Selbig, W. R., Westenbroek, S. M., & Regan, R. S. (2013). Simulation of climate-change effects on streamflow, lake water budgets, and stream temperature using GSFLOW and SNTEMP, Trout Lake Watershed, Wisconsin. USGS Scientific Investigations Report No. 2013–5159. Reston, VA: U.S. Geological Survey.

Hutton, C., Wagener, T., Freer, J., Han, D., Duffy, C., & Arheimer, B. (2016). Most computational hydrology is not reproducible, so is it really science? *Water Resources Research*, *52*(10), 7548–7555. https://doi.org/10.1002/2016WR019285

Jasechko, S., Birks, S.J., Gleeson, T., Wada, Y., Sharp, Z.D., Fawcett, P.J., McDonnell, J.J., Welker, J.M. (2014) Pronounced seasonality in the global groundwater recharge. Water Resources Research. 50, 8845–8867 doi: 10.1002/2014WR015809

Jasechko, S., Perrone, D., Befus, K. M., Bayani Cardenas, M., Ferguson, G., Gleeson, T., et al. (2017). Global aquifers dominated by fossil groundwaters but wells vulnerable to modern contamination. *Nature Geoscience*, *10*(6), 425–429. https://doi.org/10.1038/ngeo2943

Jung, M., et al. (2011). Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heatderived from eddy covariance, satellite, and meteorological observations. J. Geophys. Res.,116, G00J07,doi:10.1029/2010JG001566.

Keune, J., Sulis, M., Kollet, S., Siebert, S., & Wada, Y. (n.d.). Human Water Use Impacts on the Strength of the Continental Sink for Atmospheric Water. *Geophysical Research Letters*, *45*(9), 4068–4076. https://doi.org/10.1029/2018GL077621

Keune, J., F. Gasper, K. Goergen, A. Hense, P. Shrestha, M. Sulis, and S. Kollet, 2016, Studying the influence of groundwater representations on land surface-atmosphere feedbacks during the European heat wave in 2003, J. Geophys. Res. Atmos., 121, 13, 301–13,325, doi:10.1002/2016JD025426. doi:10.1002/2016JD025426.

Knowling, M. J., and A. D. Werner (2016), Estimability of recharge through groundwater model calibration: Insights from a field-scale steady-state example, *Journal of Hydrology*, *540*, 973-987.

Koirala et al. (2013) Global-scale land surface hydrologic modeling with the representation of water table dynamics, JGR Atmospheres https://doi.org/10.1002/2013JD020398

Koirala, S., Kim, H., Hirabayashi, Y., Kanae, S. and Oki, T. (2019) Sensitivity of Global Hydrological Simulations to Groundwater Capillary Flux Parameterizations, Water Resour. Res., 55(1), 402–425, doi:10.1029/2018WR023434,

Kollet, S. J., & Maxwell, R. M. (2008). Capturing the influence of groundwater dynamics on land surface processes using an integrated, distributed watershed model. *Water Resources Research*, *44*(2).

Kollet, S., Sulis, M., Maxwell, R. M., Paniconi, C., Putti, M., Bertoldi, G., et al. (2017). The integrated hydrologic model intercomparison project, IH-MIP2: A second set of benchmark results to diagnose integrated hydrology and feedbacks. *Water Resources Research*, *53*(1), 867–890.

Konikow, L. F. (2011), Contribution of global groundwater depletion since 1900 to sea-level rise, Geophys. Res. Lett., 38, L17401, doi: 10.1029/2011GL048604.

Koster, R.D., Suarez, M.J., Ducharne, A., Praveen, K., & Stieglitz, M. (2000). A catchment-based approach to modeling land surface processes in a GCM - Part 1: Model structure. *Journal of Geophysical Research, 105 (D20)*, 24809-24822.

Konikow, L.F. (2011) Contribution of global groundwater depletion since 1900 to sea-level rise. Geophysical Research Letters https://doi.org/10.1029/2011GL048604

Krakauer, N. Y., Li, H., & Fan, Y. (2014). Groundwater flow across spatial scales: importance for climate modeling. *Environmental Research Letters*, *9*(3), 034003.

Kresic, N. (2009). *Groundwater resources: sustainability, management and restoration*. McGraw-Hill.

Krueger, T., T. Page, K. Hubacek, L. Smith, and K. Hiscock (2012), The role of expert opinion in environmental modelling, *Environmental Modelling & Software*, *36*, 4-18.

Kustu, M. D., Fan, Y., & Rodell, M. (2011). Possible link between irrigation in the US High Plains and increased summer streamflow in the Midwest. Water Resources Research, 47(3).

Lamb, R., Aspinall, W., Odbert, H. and Wagener, T. (2017). Vulnerability of bridges to scour: Insights from an international expert elicitation workshop. Natural Hazards and Earth System Sciences. 17(8), 1393-1409.

Leaf, A. T., Fienen, M. N., Hunt, R. J., & Buchwald, C. A. (2015). Groundwater/surface-water interactions in the Bad River Watershed, Wisconsin. USGS Numbered Series No. 2015–5162. Reston, VA: U.S. Geological Survey.

Leavesley, G. H., S. L. Markstrom, P. J. Restrepo, and R. J. Viger (2002), A modular approach for addressing model design, scale, and parameter estimation issues in distributed hydrological modeling, Hydrol. Processes, 16, 173–187, doi:10.1002/hyp.344.

Lemieux, J. M., Sudicky, E. A., Peltier, W. R., & Tarasov, L. (2008). Dynamics of groundwater recharge and seepage over the Canadian landscape during the Wisconsinian glaciation. *J. Geophys. Res.*, *113*. Retrieved from http://dx.doi.org/10.1029/2007JF000838

Lenton, T.M. et al. (2008). Tipping elements in the Earth's climate system. Proceedings of the National Academy of Sciences 105 (6), 1786-1793.

Liang, X., Z. Xie, and M. Huang (2003). A new parameterization for surface and groundwater interactions and its impact on water budgets with the variable infiltration capacity (VIC) land surface model, *J. Geophys. Res.*, 108, 8613, D16. https://doi.org/10.1029/2002JD003090

Lo, M.-H., Famiglietti, J. S., Reager, J. T., Rodell, M., Swenson, S., & Wu, W.-Y. (2016). GRACE-Based Estimates of Global Groundwater Depletion. In Q. Tang & T. Oki (Eds.), *Terrestrial Water Cycle and Climate Change* (pp. 135–146). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118971772.ch7

Lo, M.-H., Yeh, P. J.-F., & Famiglietti, J. S. (2008). Constraining water table depth simulations in a land surface model using estimated baseflow. *Advances in Water Resources*, *31*(12), 1552–1564.

Lo, M. and J. S. Famiglietti, (2010) Effect of water table dynamics on land surface hydrologic memory, J. Geophys. Res., 115, D22118, doi:10.1029/2010JD014191

Lo, M.-H., J. S. Famiglietti, P. J.-F. Yeh, and T. H. Syed (2010), Improving Parameter Estimation and Water Table Depth Simulation in a Land Surface Model Using GRACE Water Storage and Estimated Baseflow Data, Water Resour. Res., 46, W05517, doi:10.1029/2009WR007855.

Loheide, S. P., Butler Jr, J. J., & Gorelick, S. M. (2005). Estimation of groundwater consumption by phreatophytes using diurnal water table fluctuations: A saturated-unsaturated flow assessment. *Water Resources Research*, *41*(7).

Luijendijk, E., Gleeson, T. and Moosdorf, N. (2020) Fresh groundwater discharge insignificant for the world's oceans but important for coastal ecosystems Nature Communications, 11, 1260 (2020). doi: 10.1038/s41467-020-15064-8

Maples, S., Foglia, L., Fogg, G.E. and Maxwell, R.M. (2020). Sensitivity of Hydrologic and Geologic Parameters on Recharge Processes in a Highly-Heterogeneous, Semi-Confined Aquifer System. Hydrology and Earth Systems Sciences, in press.

Margat, J., & Van der Gun, J. (2013). *Groundwater around the world: a geographic synopsis*. London: CRC Press

Markovich, KH, AH Manning, LE Condon, JC McIntosh (2019). Mountain-block Recharge: A Review of Current Understanding. Water Resources Research, 55, https://doi.org/10.1029/2019WR025676

Maxwell, R. M., Condon, L. E., and Kollet, S. J. (2015) A high-resolution simulation of groundwater and surface water over most of the continental US with the integrated hydrologic model ParFlow v3, Geosci. Model Dev., 8, 923–937, https://doi.org/10.5194/gmd-8-923-2015.

Maxwell, R.M., Chow, F.K. and Kollet, S.J., The groundwater-land-surface-atmosphere connection: soil moisture effects on the atmospheric boundary layer in fully-coupled simulations. Advances in Water Resources 30(12), doi:10.1016/j.advwatres.2007.05.018, 2007.

Maxwell, R. M., & Condon, L. E. (2016). Connections between groundwater flow and transpiration partitioning. *Science*, *353*(6297), 377–380.

Maxwell, R. M., Condon, L. E., Kollet, S. J., Maher, K., Haggerty, R., & Forrester, M. M. (2016). The imprint of climate and geology on the residence times of groundwater. *Geophysical Research Letters*, *43*(2), 701–708. https://doi.org/10.1002/2015GL066916

McMilan, H. (2020). Linking hydrologic signatures to hydrologic processes: A review. Hydrological Processes. 34, 1393– 1409.

Meixner, T., Manning, A. H., Stonestrom, D. A., Allen, D. M., Ajami, H., Blasch, K. W., et al. (2016). Implications of projected climate change for groundwater recharge in the western United States. *Journal of Hydrology*, *534*, 124–138.

Melsen, L. A., A. J. Teuling, P. J. J. F. Torfs, R. Uijlenhoet, N. Mizukami, and M. P. Clark, 2016a: HESS Opinions: The need for process-based evaluation of large-domain hyper-resolution models. Hydrology and Earth System Sciences, doi:10.5194/hess-20-1069-2016.

Meriano, M., & Eyles, N. (2003). Groundwater flow through Pleistocene glacial deposits in the rapidly urbanizing Rouge River-Highland Creek watershed, City of Scarborough, southern Ontario, Canada. Hydrogeology Journal, 11(2), 288–303. https://doi.org/10.1007/s10040-002-0226-4

Milly, P.C., S.L. Malyshev, E. Shevliakova, K.A. Dunne, K.L. Findell, T. Gleeson, Z. Liang, P. Phillipps, R.J. Stouffer, & S. Swenson (2014). An Enhanced Model of Land Water and Energy for Global Hydrologic and Earth-System Studies. *J. Hydrometeor.,* 15, 1739–1761. https://doi.org/10.1175/JHM-D-13-0162.1

Minderhoud P S J, Erkens G, Pham Van H, Bui Tran V, Erban L E, Kooi, H and Stouthamer E (2017) Impacts of 25 years of groundwater extraction on subsidence in the Mekong delta, Vietnam Environ. Res. Lett. 12 064006

Minderhoud, P.S.J., Coumou, L., Erkens, G., Middelkoop, H. & Stouthamer, E. (2019). Mekong delta much lower than previously assumed in sea-level rise impact assessments. Nature Communications 10, 3847.

Minderhoud, P.S.J., Middelkoop, H., Erkens, G. and Stouthamer, E. Groundwater (2020). extraction may drown mega-delta: projections of extraction-induced subsidence and elevation of the Mekong delta for the 21st century. Environ. Res. Commun. 2, 011005.

Miralles, D. G., Jimenez, C., Jung, M., Michel, D., Ershadi, A., McCabe, M. F., et al. (2016). The WACMOS-ET project - Part 2: Evaluation of global terrestrial evaporation data sets. *Hydrology and Earth System Sciences, 20*(2), 823-842. doi:10.5194/hess-20-823-2016.

Moeck, C. Nicolas Grech-Cumbo, Joel Podgorski, Anja Bretzler, Jason J. Gurdak ,Michael Berg, Mario Schirmer (2020) A global-scale dataset of direct natural groundwater recharge rates: A review of variables, processes and relationships. Science of The Total Environment https://doi.org/10.1016/j.scitotenv.2020.137042

Mohan, C., Wei, Y., & Saft, M. (2018). Predicting groundwater recharge for varying land cover and climate conditions–a global meta-study. *Hydrology and Earth System Sciences*, *22*(5), 2689–2703.

Montanari, A., Young, G., Savenije, H.H.G., Hughes, D., Wagener, T., Ren, L.L., Koutsoyiannis, D., Cudennec, C., Toth, E., Grimaldi, S., et al. (2013). "Panta Rhei—Everything Flows": Change in hydrology and society—The IAHS Scientific Decade 2013–2022. Hydrological Sciences Journal *58*, 1256–1275.

Moore, W. S. (2010). The effect of submarine groundwater discharge on the ocean. *Annual Review of Marine Science*, *2*, 59–88.

Morris, M. D. (1991). Factorial sampling plans for preliminary computational experiments. *Technometrics*, *33*(2), 161–174.

Müller Schmied, H., Eisner, S., Franz, D., Wattenbach, M., Portmann, F.T., Flörke, M., Döll, P. (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, 3511-3538, doi: 10.5194/hess-18-3511-2014.

Niu, G.-Y., Z.-L. Yang, R. E. Dickinson, and L. E. Gulden (2005), A simple TOPMODEL-based runoff parameterization (SIMTOP) for use in global climate models. *J. Geophys. Res.*, 110, D21106, doi:10.1029/2005JD006111

Niu GY, Yang ZL, Dickinson RE, Gulden LE, Su H (2007) Development of a simple groundwater model for use in climate models and evaluation with Gravity Recovery and Climate Experiment data. J Geophys Res 112:D07103. doi:10.1029/2006JD007522

Ngo-Duc, T., Laval, K. Ramillien, G., Polcher, J. & Cazenave, A. (2007). Validation of the land water storage simulated by Organising Carbon and Hydrology in Dynamic Ecosystems (ORCHIDEE) with Gravity Recovery and Climate Experiment (GRACE) data. *Water Resour. Res.*, 43, W04427. https://doi.org/10.1029/2006WR004941

O'Hagan, A. (2019). Expert Knowledge Elicitation: Subjective but Scientific. The American Statistician, 73, doi.org/10.1080/00031305.2018.1518265

Olarinoye, T., et al. (2020): Global karst springs hydrograph dataset for research and management of the world's fastest-flowing groundwater, Sci. Data, 7(1), doi:10.1038/s41597-019-0346-5.

Opie, S., Taylor, R. G., Brierley, C. M., Shamsudduha, M., & Cuthbert, M. O. (2020). Climate–groundwater dynamics inferred from GRACE and the role of hydraulic memory. *Earth System Dynamics*, *11*(3), 775–791. https://doi.org/10.5194/esd-11-775-2020

Ortega-Guerrero A, Rudolph D L and Cherry J A 1999 Analysis of long-term land subsidence near Mexico City: field investigations and predictive modeling Water Resour. Res. 353327–41

Pan, M., Sahoo, A. K., Troy, T. J., Vinukollu, R. K., Sheffield, J., & Wood, F. E. (2012). Multisource estimation of long-term terrestrial water budget for major global river basins. *J. Climate*, 25, 3191–3206. https://doi.org/10.1175/JCLI-D-11-00300.1

Pappenberger, F., Ghelli, A., Buizza, R. and Bodis, K. (2009). The Skill of Probabilistic Precipitation Forecasts under Observational Uncertainties within the Generalized Likelihood Uncertainty Estimation Framework for Hydrological Applications. Journal of Hydrometeorology, DOI: 10.1175/2008JHM956.1

Pellet, V., Aires, F., Munier, S., Fernández Prieto, D., Jordá, G., Dorigo, W. A., Polcher, J., & Brocca, L. (2019). Integrating multiple satellite observations into a coherent dataset to monitor the full water cycle – application to the Mediterranean region. *Hydrol. Earth Syst. Sci.*, 23, 465-491. https://doi.org/10.5194/hess-23-465-2019

Perrone, D. and Jasechko (2019). Deeper well drilling an unsustainable stopgap to groundwater depletion. *Nature Sustain.* 2, 773-782.

Person, M. A., Raffensperger, J. P., Ge, S., & Garven, G. (1996). Basin-scale hydrogeologic modeling. *Reviews of Geophysics*, *34*(1), 61–87.

Pianosi, F., Beven, K., Freer, J., Hall, J. W., Rougier, J., Stephenson, D. B., & Wagener, T. (2016). Sensitivity analysis of environmental models: A systematic review with practical workflow. *Environmental Modelling & Software*, *79*, 214–232.

Post, V. E., & von Asmuth, J. R. (2013). Hydraulic head measurements—new technologies, classic pitfalls. *Hydrogeology Journal*, *21*(4), 737–750.

Qiu J. Q., Zipper, S.C., Motew M., Booth, E.G., Kucharik, C.J., & Loheide, S.P. (2019). Nonlinear groundwater influence on biophysical indicators of ecosystem services. *Nature Sustainability*, in press, doi: 10.1038/s41893-019-0278-2

Rajabi, M. M., and B. Ataie-Ashtiani (2016), Efficient fuzzy Bayesian inference algorithms for incorporating expert knowledge in parameter estimation, *Journal of Hydrology*, *536*, 255-272.

Rajabi, M. M., B. Ataie-Ashtiani, and C. T. Simmons (2018), Model-data interaction in groundwater studies: Review of methods, applications and future directions, *Journal of Hydrology*, *567*, 457-477.

Rashid, M., Chien, R.Y., Ducharne, A., Kim, H., Yeh, P.J.F., Peugeot, C., Boone, A., He, X., Séguis, L., Yabu, Y., Boukari, M. & Lo, M.H. (2019). Evaluation of groundwater simulations in Benin from the ALMIP2 project. *J. Hydromet.*, accepted.

Refsgaard, J.C., van der Sluijs, J.P., Højberg, A.L., and Vanrolleghem, P.A. (2007). Uncertainty in the environmental modelling process–a framework and guidance. Environmental Modelling & Software, 22(11), 1543-1556

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204.

Reinecke, R., Foglia, L., Mehl, S., Trautmann, T., Cáceres, D., & Döll, P. (2019a). Challenges in developing a global gradient-based groundwater model. (G^{3}M v1.0) for the integration into a global hydrological model. Geosci. Model Dev., 12, 2401-2418. doi: 10.5194/gmd-12-2401-2019

Reinecke, R., Foglia, L., Mehl, S., Herman, J., Wachholz, A., Trautmann, T., and Döll, P. (2019b) Spatially distributed sensitivity of simulated global groundwater heads and flows to hydraulic conductivity, groundwater recharge and surface water body parameterization, Hydrology and Earth System Sciences, (23) 4561–4582. 2019.

Reinecke, R., Wachholz, A., Mehl, S., Foglia, L., Niemann, C., Döll, P. (2020). Importance of spatial resolution in global groundwater modeling. Groundwater. doi: 10.1111/gwat.12996

Rodell, M., Velicogna, I., & Famiglietti, J. S. (2009). Satellite-based estimates of groundwater depletion in India. *Nature*, *460*(7258), 999–1002.

Rodell, M., Famiglietti, J. S., Wiese, D. N., Reager, J. T., Beaudoing, H. K., Landerer, F. W., & Lo, M.-H. (2018). Emerging trends in global freshwater availability. *Nature*, 557(7707), 651.

Rosolem, R., Hoar, T., Arellano, A., Anderson, J. L., Shuttleworth, W. J., Zeng, X., and Franz, T. E.: Translating aboveground cosmic-ray neutron intensity to high-frequency soil moisture profiles at sub-kilometer scale, Hydrol. Earth Syst. Sci., 18, 4363-4379

Ross, J. L., M. M. Ozbek, and G. F. Pinder (2009), Aleatoric and epistemic uncertainty in groundwater flow and transport simulation, *Water Resources Research*, *45*(12).

Rossman, N., & Zlotnik, V. (2013). Review: Regional groundwater flow modeling in heavily irrigated basins of selected states in the western United States. *Hydrogeology Journal*, *21*(6), 1173–1192. https://doi.org/10.1007/s10040-013-1010-3

RRCA. (2003). Republican River Compact Administration Ground Water Model. Retrieved from http://www.republicanrivercompact.org/

Saltelli, A., Chan, K., & Scott, E. M. (Eds.). (2000). *Sensitivity analysis*. Wiley.

Salvucci, G. D., & Entekhabi, D. (1995). Hillslope and climatic controls on hydrologic fluxes. *Water Resources Research*, *31*(7), 1725–1739.

Sanford, W. Calibration of models using groundwater age. Hydrogeol J 19, 13–16 (2011). https://doi.org/10.1007/s10040-010-0637-6

Sawyer, A. H., David, C. H., & Famiglietti, J. S. (2016). Continental patterns of submarine groundwater discharge reveal coastal vulnerabilities. *Science*, *353*(6300), 705–707.

Scanlon, B., Healy, R., & Cook, P. (2002). Choosing appropriate techniques for quantifying groundwater recharge. *Hydrogeology Journal*, *10*(1), 18–39.

Scanlon, B. R., Keese, K. E., Flint, A. L., Flint, L. E., Gaye, C. B., Edmunds, W. M., & Simmers, I. (2006). Global synthesis of groundwater recharge in semiarid and arid regions. *Hydrological Processes*, *20*, 3335–3370.

Scanlon, B. R., Faunt, C. C., Longuevergne, L., Reedy, R. C., Alley, W. M., McGuire, V. L., & McMahon, P. B. (2012). Groundwater depletion and sustainability of irrigation in the US High Plains and Central Valley. *Proceedings of the National Academy of Sciences*, *109*(24), 9320–9325. https://doi.org/10.1073/pnas.1200311109

Scanlon, B. R., Zhang, Z., Save, H., Wiese, D. N., Landerer, F. W., Long, D., et al. (2016). Global evaluation of new GRACE mascon products for hydrologic applications. *Water Resources Research*, *52*(12), 9412–9429.

Scanlon, B. R., Zhang, Z., Save, H., Sun, A. Y., Müller Schmied, H., van Beek, L. P., et al. (2018). Global models underestimate large decadal declining and rising water storage trends relative to GRACE satellite data. *Proceedings of the National Academy of Sciences*, 201704665.

Schaller, M., and Y. Fan (2009) River basins as groundwater exporters and importers: Implications for water cycle and climate modeling. Journal of Geophysical Research-Atm, 114, D04103, doi: 10.1029/2008 JD010636

Schewe, J., Heinke, J., Gerten, D., Haddeland, I., Arnell, N. W., Clark, D. B., et al. (2014). Multimodel assessment of water scarcity under climate change. *Proceedings of the National Academy of Sciences*, *111*(9), 3245–3250. https://doi.org/10.1073/pnas.1222460110

Schilling, O. S., Doherty, J., Kinzelbach, W., Wang, H., Yang, P. N., & Brunner, P. (2014). Using tree ring data as a proxy for transpiration to reduce predictive uncertainty of a model simulating groundwater–surface water–vegetation interactions. Journal of Hydrology, 519, Part B, 2258–2271. https://doi.org/10.1016/j.jhydrol.2014.08.063

Schilling, O.S., Cook, P.G., Brunner, P., 2019. Beyond classical observations in hydrogeology: The advantages of including exchange flux, temperature, tracer concentration, residence time, and soil moisture observations in groundwater model calibration. Reviews of Geophysics, 57(1): 146-182.

Schneider, A.S., Jost, A., Coulon, C., Silvestre, M., Théry, S., & Ducharne, A. (2017). Global scale river network extraction based on high-resolution topography, constrained by lithology, climate, slope, and observed drainage density. *Geophysical Research Letters*, 44, 2773–2781. https://doi.org/10.1002/2016GL071844

Shen, C. (2018). A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research*, *54*(11), 8558–8593.

Shen, C., Laloy, E., Elshorbagy, A., Albert, A., Bales, J., Chang, F.-J., et al. (2018). HESS Opinions: Incubating deep-learning-powered hydrologic science advances as a community. *Hydrology and Earth System Sciences*, *22*(11).

SKI (1984). Intracoin - International Nuclide Transport Code Intercomparison Study (No. SKI--84-3). Swedish Nuclear Power Inspectorate. Retrieved from http://inis.iaea.org/Search/search.aspx?orig_q=RN:16046803

Springer, A., & Stevens, L. (2009). Spheres of discharge of springs. Hydrogeology Journal, 17(1), 83–93. https://doi.org/10.1007/s10040-008-0341-y

Steffen, W., Broadgate, W., Deutsch, L., Gaffney, O., & Ludwig, C. (2015). The trajectory of the Anthropocene: the great acceleration. *The Anthropocene Review*, *2*(1), 81–98.

Sutanudjaja, E. H., Beek, R. van, Wanders, N., Wada, Y., Bosmans, J. H., Drost, N., et al. (2018). PCR-GLOBWB 2: a 5 arcmin global hydrological and water resources model. *Geoscientific Model Development*, *11*(6), 2429–2453.

Takata, K., Emori, S. and Watanabe, T.: Development of the minimal advanced treatments of surface interaction and runoff, Glob. Planet. Change, 38(1–2), 209–222, doi:10.1016/S0921-8181(03)00030-4, 2003.

Tallaksen, L. M. (1995). A review of baseflow recession analysis. *Journal of Hydrology*, *165*(1–4), 349–370. https://doi.org/10.1016/0022-1694(94)02540-R

Taylor, R. G., Todd, M. C., Kongola, L., Maurice, L., Nahozya, E., Sanga, H., & MacDonald, A. M. (2013b). Evidence of the dependence of groundwater resources on extreme rainfall in East Africa. *Nature Clim. Change*, *3*(4), 374–378. https://doi.org/10.1038/nclimate1731

Taylor, R. G., Scanlon, B., Doll, P., Rodell, M., van Beek, R., Wada, Y., et al. (2013a). Groundwater and climate change. *Nature Clim. Change*, *3*(4), 322–329. https://doi.org/10.1038/nclimate1744

Thatch, L. M., Gilbert, J. M., & Maxwell, R. M. (2020). Integrated hydrologic modeling to untangle the impacts of water management during drought. *Groundwater*, *58*(3), 377–391.

Thomas, Z., Rousseau-Gueutin, P., Kolbe, T., Abbott, B.W., Marçais, J., Peiffer, S., Frei, S., Bishop, K., Pichelin, P., Pinay, G., de Dreuzy, J.R. (2016). Constitution of a catchment virtual observatory for sharing flow and transport models outputs. *Journal of Hydrology*, 543, Pages 59-66. https://doi.org/10.1016/j.jhydrol.2016.04.067

Tolley, D., Foglia, L., & Harter, T. (2019). Sensitivity Analysis and Calibration of an Integrated Hydrologic Model in an Irrigated Agricultural Basin with a Groundwater-Dependent Ecosystem. Water Resources Research. https://doi.org/10.1029/2018WR024209

Tóth, J. (1963). A theoretical analysis of groundwater flow in small drainage basins. *Journal of Geophysical Research*, *68*(16), 4795–4812.

Tran, H., Jun Zhang, Jean-Martial Cohard, Laura E. Condon, Reed M. Maxwell (20200) Simulating g roundwater-Streamflow Connections in the Upper Colorado River Basin Groundwater, 2020 https://doi.org/10.1111/gwat.13000

Tregoning, P., McClusky, S., van Dijk, A.I.J.M. and Crosbie, R.S. (2012). Assessment of GRACE satellites for groundwater estimation in Australia. Waterlines Report Series No 71, National Water Commission, Canberra

Troldborg, L., Refsgaard, J. C., Jensen, K. H., & Engesgaard, P. (2007). The importance of alternative conceptual models for simulation of concentrations in a multi-aquifer system. Hydrogeology Journal, 15(5), 843–860.

Tustison, B., Harris, D. and Foufoula-Georgiou, E. (2001). Scale issues in verification of precipitation forecasts. Journal of geophysical Research, 106(D11), 11775-11784.

UNESCO. (1978). *World water balance and water resources of the earth* (Vol. USSR committee for the international hydrologic decade). Paris: UNESCO.

van Vliet, M. T., Flörke, M., Harrison, J. A., Hofstra, N., Keller, V., Ludwig, F., et al. (2019). Model inter-comparison design for large-scale water quality models. Current Opinion in Environmental Sustainability, 36, 59–67. https://doi.org/10.1016/j.cosust.2018.10.013

Van Werkhoven, K., Wagener, T., Tang, Y., and Reed, P. 2008. Understanding watershed model behavior across hydro-climatic gradients using global sensitivity analysis. Water Resources Research, 44, W01429, doi:10.1029/2007WR006271.

Van Loon, A.F. et al. (2016) Drought in the Anthropocene. *Nature Geoscience* 9: 89-91 doi: 10.1038/ngeo2646.

van Loon, Anne F.; Kumar, Rohini; Mishra, Vimal (2017): Testing the use of standardised indices and GRACE satellite data to estimate the European 2015 groundwater drought in near-real time. In Hydrol. Earth Syst. Sci. 21 (4), pp. 1947–1971. DOI: 10.5194/hess-21-1947-2017.

Vergnes, J.-P., & Decharme, B. (2012). A simple groundwater scheme in the TRIP river routing model: global off-line evaluation against GRACE terrestrial water storage estimates and observed river discharges. *Hydrol. Earth Syst. Sci.*, 16, 3889-3908. https://doi.org/10.5194/hess-16-3889-2012

Vergnes, J.-P., B. Decharme, & F. Habets (2014). Introduction of groundwater capillary rises using subgrid spatial variability of topography into the ISBA land surface model, *J. Geophys. Res. Atmos.*, 119, 11,065–11,086. https//doi.org/10.1002/2014JD021573

Vergnes, J.-P., Roux, N., Habets, F., Ackerer, P., Amraoui, N., Besson, F., et al. (2020). The AquiFR hydrometeorological modelling platform as a tool for improving groundwater resource monitoring over France: evaluation over a 60-year period. Hydrology and Earth System Sciences, 24(2), 633–654. https://doi.org/10.5194/hess-24-633-2020

Visser, W. C. (1959). Crop growth and availability of moisture. *Journal of the Science of Food and Agriculture*, *10*(1), 1–11.

Wada, Y., L. P. H. van Beek, C. M. van Kempen, J. W. T. M. Reckman, S. Vasak, M. F. P. Bierkens, (2010) Global depletion of groundwater resources. Geophys. Res. Lett. 37, L20402.

Wada, Y.; Wisser, D.; Bierkens, M. F. P. (2014).Global modeling of withdrawal, allocation and consumptive use of surface water and groundwater resources. Earth System Dynamics Discussions, volume 5, issue 1, pp. 15 - 40

Wada, Y. (2016). Modeling Groundwater Depletion at Regional and Global Scales: Present State and Future Prospects. *Surveys in Geophysics*, *37*(2), 419–451. https://doi.org/10.1007/s10712-015-9347-x

Wada, Y., & Bierkens, M. F. P. (2014). Sustainability of global water use: past reconstruction and future projections. Environmental Research Letters, 9(10), 104003. https://doi.org/10.1088/1748-9326/9/10/104003

Wada, Y., & Heinrich, L. (2013). Assessment of transboundary aquifers of the world—vulnerability arising from human water use. *Environmental Research Letters*, *8*(2), 024003.

Wagener, T. 2003. Evaluation of catchment models. Hydrological Processes, 17, 3375-3378.

Wagener, T., & Gupta, H. V. (2005). Model identification for hydrological forecasting under uncertainty. *Stochastic Environmental Research and Risk Assessment*, *19*(6), 378–387.

Wagener, T., Sivapalan, M., Troch, P. and Woods, R. (2007). Catchment classification and hydrologic similarity. Geography Compass, 1(4), 901, doi:10.1111/j.1749-8198.2007.00039.x

Wagener, T. and Pianosi, F. (2019) What has Global Sensitivity Analysis ever done for us? A systematic review to support scientific advancement and to inform policy-making in earth system modelling. Earth-Science Reviews, 194, 1-18. doi.org/10.1016/j.earscirev.2019.04.006

Wagener, T., Boyle, D.P., Lees, M.J., Wheater, H.S., Gupta, H.V. and Sorooshian, S. (2001). A framework for development and application of hydrological models. Hydrology and Earth System Sciences, 5(1), 13-26.

Wagener, T., Sivapalan, M., Troch, P. A., McGlynn, B. L., Harman, C. J., Gupta, H. V., et al. (2010). The future of hydrology: An evolving science for a changing world. *Water Resources Research*, *46*(5).

Wagener, T., Gleeson, T., Coxon, G., Hartmann, A., Howden, N., Pianosi, F., Rahman, M., Rosolem, R., Stein, L., and Woods, R. (2021). On doing hydrology with dragons: Realizing the value of perceptual models and knowledge accumulation. *Wiley Interdisciplinary Reviews: Water*, e1550. https://doi.org/10.1002/wat2.1550

Wang, F., Ducharne, A., Cheruy, F., Lo, M.H., & Grandpeix, J.L. (2018). Impact of a shallow groundwater table on the global water cycle in the IPSL land-atmosphere coupled model, *Climate Dynamics*, *50*, 3505-3522, https://doi.org/10.1007/s00382-017-3820-9

Warszawski, L., Frieler, K., Huber, V., Piontek, F., Serdeczny, O., & Schewe, J. (2014). The Inter-Sectoral Impact Model Intercomparison Project (ISI–MIP): Project framework. *Proceedings of the National Academy of Sciences*, *111*(9), 3228–3232. https://doi.org/10.1073/pnas.1312330110

Winter, T. C., Harvey, J. W., Franke, O. L., & Alley, W. M. (1998). *Ground water and surface water: a single resource* (p. 79). U.S. Geological Survey circular 1139

Woolfenden, L. R., & Nishikawa, T. (2014). Simulation of groundwater and surface-water resources of the Santa Rosa Plain watershed, Sonoma County, California. USGS Scientific Investigations Report 2014–5052). Reston, VA: U.S. Geological Survey.

Yang, J., Griffiths, J., & Zammit, C. (2019). National classification of surface–groundwater interaction using random forest machine learning technique. *River Research and Applications*, *35*(7), 932–943. https://doi.org/10.1002/rra.3449

Yeh, P. J.-F. and J. Famiglietti, Regional groundwater evapotranspiration in Illinois, J. Hydrometeorology, 10(2), 464–478, 2010

Yilmaz, K., Gupta, H.V. and Wagener, T. 2009. Towards improved distributed modeling of watersheds: A process based diagnostic approach to model evaluation. Water Resources Research, 44, W09417, doi:10.1029/2007WR006716.

Young, P., Parkinson, S. and Lees, M. (1996). Simplicity out of complexity in environmental modelling: Occam's razor revisited. Journal of Applied Statistics, 23(2-3), 165-210. https://doi.org/10.1080/02664769624206

Zell, W. O., & Sanford, W. E. (2020). Calibrated Simulation of the Long-Term Average Surficial Groundwater System and Derived Spatial Distributions of its Characteristics for the Contiguous United States. Water Resources Research, 56(8), e2019WR026724. https://doi.org/10.1029/2019WR026724

Zipper, S. C., Soylu, M. E., Booth, E. G., & Loheide, S. P. (2015). Untangling the effects of shallow groundwater and soil texture as drivers of subfield-scale yield variability. *Water Resources Research*, 51(8), 6338–6358.

Zipper, S. C., Soylu, M. E., Kucharik, C. J., & Loheide, S. P. (2017). Quantifying indirect groundwater-mediated effects of urbanization on agroecosystem productivity using MODFLOW-AgroIBIS (MAGI), a complete critical zone model. *Ecological Modelling*, 359, 201-219

Zhang, M and Burbey T J 2016 Inverse modelling using PS-InSAR data for improved land subsidence simulation in Las Vegas Valley, Nevada Hydrol. Process. 30 4494–516

Zhou, Y., Li, W., 2011. A review of regional groundwater flow modeling. Geoscience Frontiers, 2(2): 205-214.