

Towards a theory of evaluation for global water models

Sebastian Gnann & Thorsten Wagener



The need for a “theory of evaluation”

- Global water models are used for high impact studies & policymaking
- We thus need adequate evaluation strategies for large scale models
- However, we do not have an agreed upon evaluation framework



Safe and just Earth system boundaries

<https://doi.org/10.1038/s41586-023-06083-8>

Received: 23 June 2022

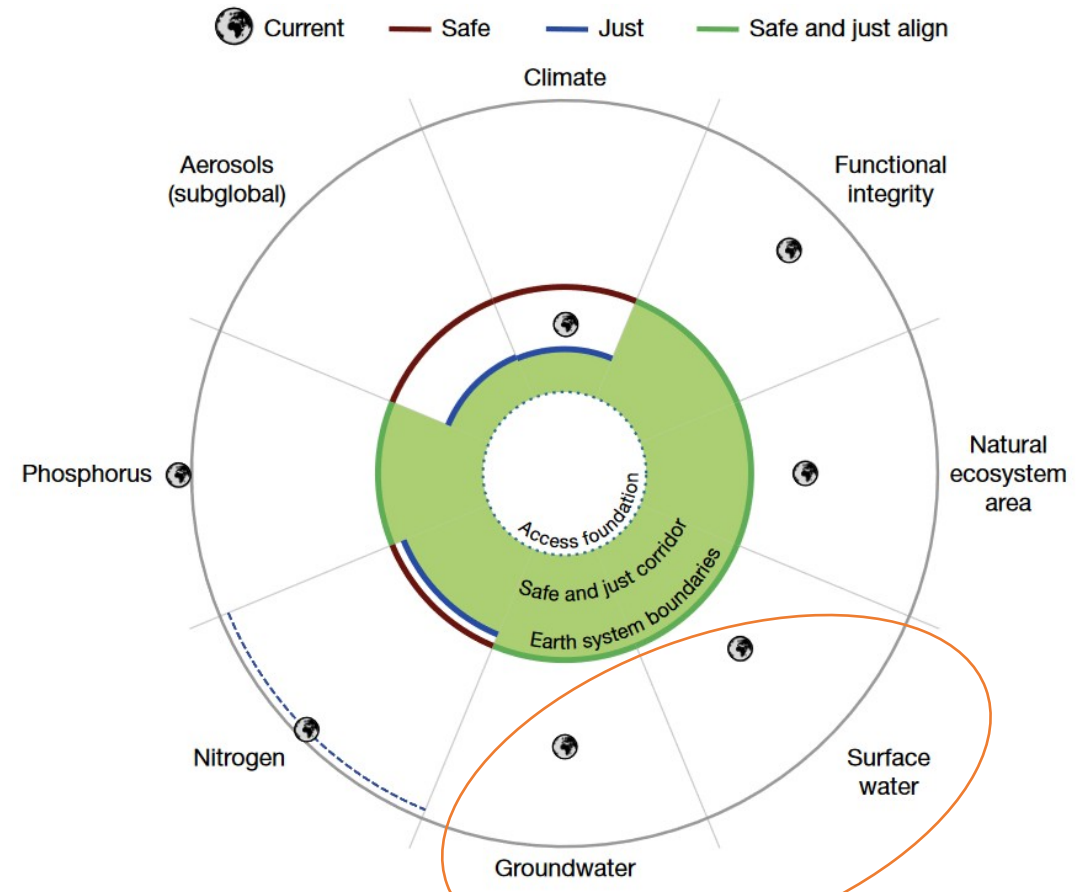
Accepted: 14 April 2023

Published online: 31 May 2023

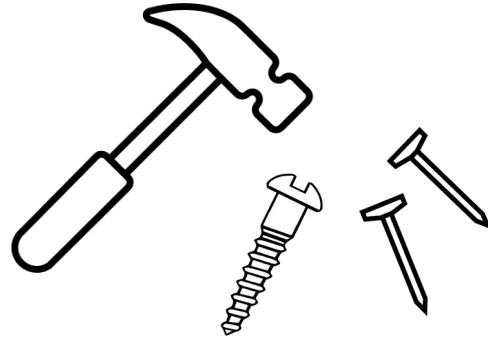
Open access

Check for updates

Johan Rockström^{1,2,3}, Joyeeta Gupta^{4,5}, Dahe Qin^{6,7,8}, Steven J. Lade^{9,10}, Jesse F. Abrams¹¹, Lauren S. Andersen¹, David I. Armstrong McKay^{3,11,12}, Xuemei Bai¹⁰, Govindasamy Bala¹³, Stuart E. Bunn¹⁴, Daniel Ciobanu³, Fabrice DeClerck^{15,16}, Kristie Ebi¹⁷, Lauren Gifford¹⁸, Christopher Gordon¹⁹, Syezlin Hasan¹⁴, Norichika Kanie²⁰, Timothy M. Lenton¹¹, Sina Loriani¹, Diana M. Liverman¹⁸, Awaz Mohamed²¹, Nebojsa Nakicenovic²², David Obura²³, Daniel Ospina⁹, Klaudia Prodani⁴, Crelis Rammelt⁴, Boris Sakschewski¹, Joeri Scholtens⁴, Ben Stewart-Koster¹⁴, Thejna Tharammal²⁴, Detlef van Vuuren^{25,26}, Peter H. Verburg^{27,28}, Ricarda Winkelmann^{1,29}, Caroline Zimm²², Elena M. Bennett^{30,31}, Stefan Brinzeu³², Wendy Broadgate⁹, Pamela A. Green³³, Lei Huang³⁴, Lisa Jacobson⁹, Christopher Ndehedehe^{14,35}, Simona Pedde^{9,36}, Juan Rocha^{3,9}, Marten Scheffer²⁷, Lena Schulte-Uebbing^{25,36}, Wim de Vries³⁶, Cunde Xiao^{6,39}, Chi Xu⁴⁰, Xinwu Xu^{7,8}, Noelia Zafra-Calvo⁴¹ & Xin Zhang⁴²



An evaluation framework should be **fit for purpose**



Uses of models

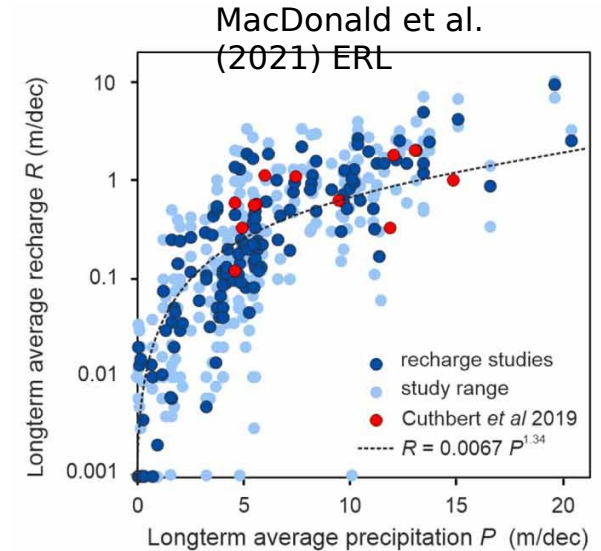
- Large scale
- Data scarce regions
- Changing conditions

Users of models

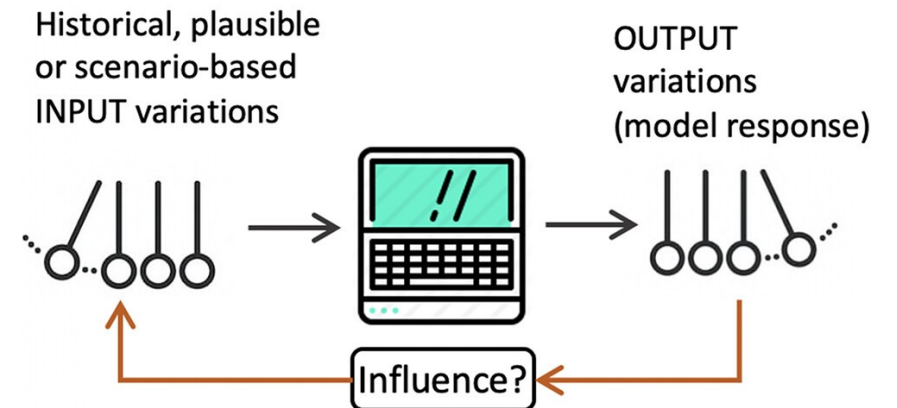
- **Researchers**
- **Model developers**
- Stakeholders

Fit for purpose evaluation for uses of models

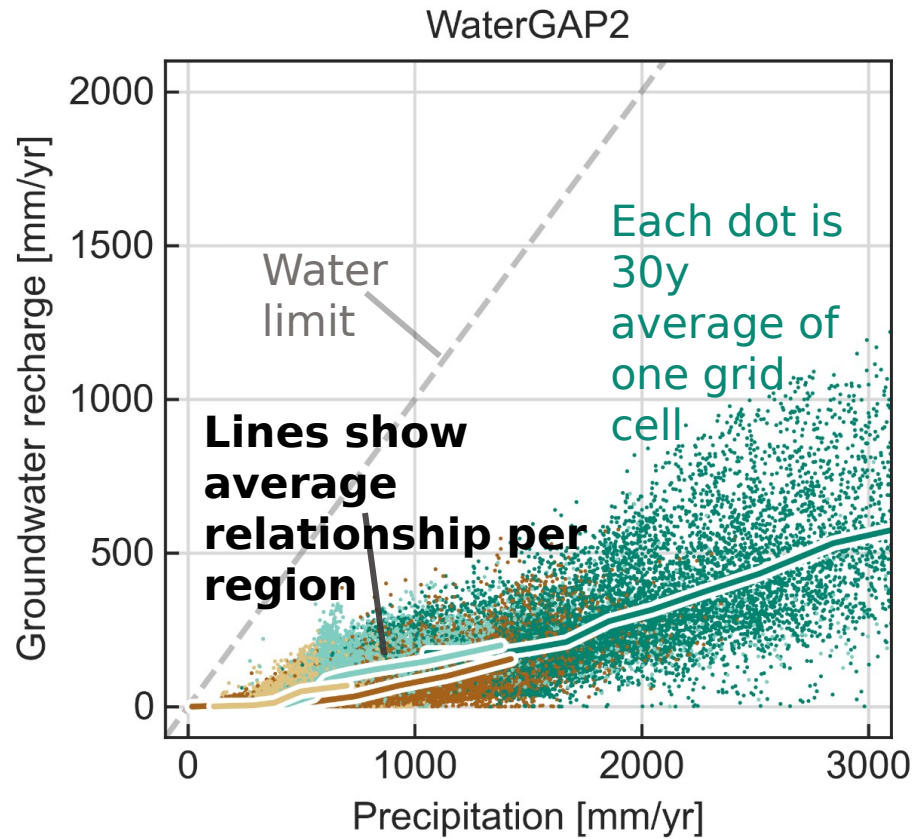
- Evaluation is often done as for catchment scale models, but there are alternatives
- Evaluation using **functional relationships** that capture large scale relationships across places
- Response-based evaluation to evaluate how models respond to changing conditions



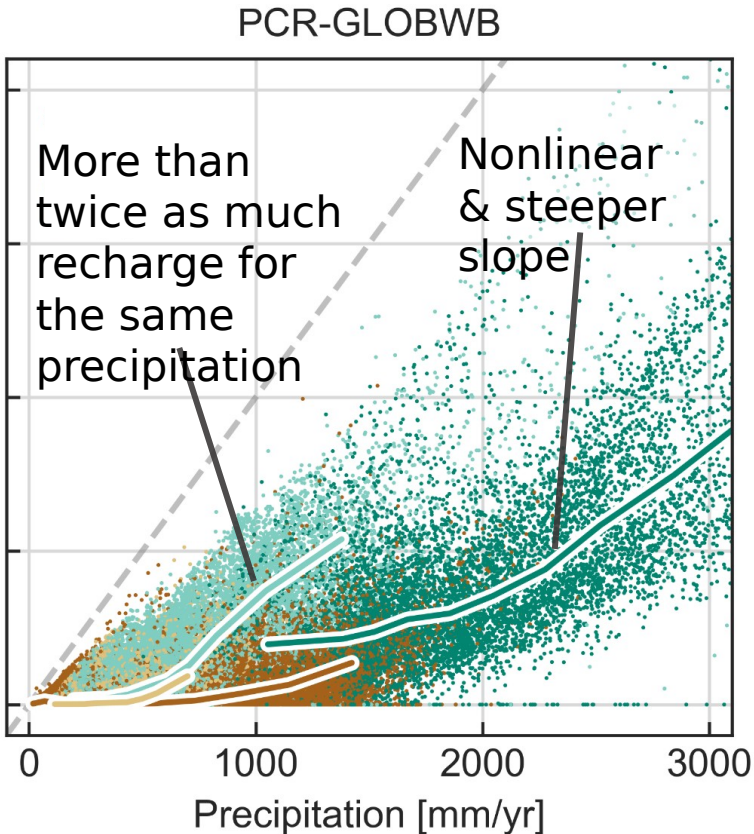
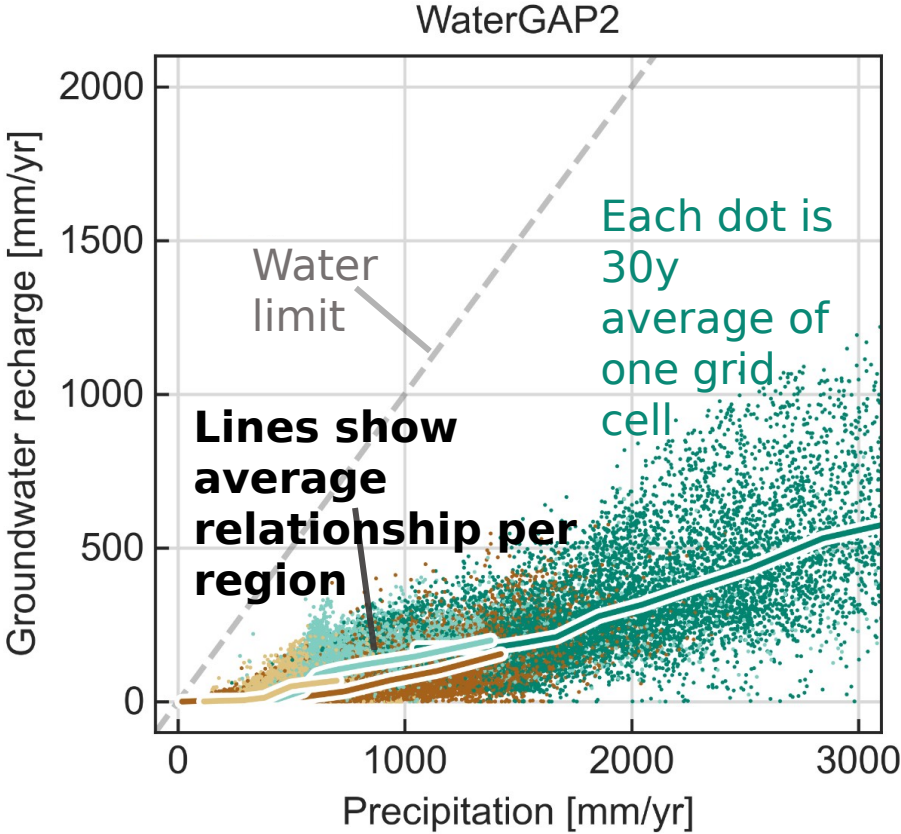
Wagener et al. (2022)



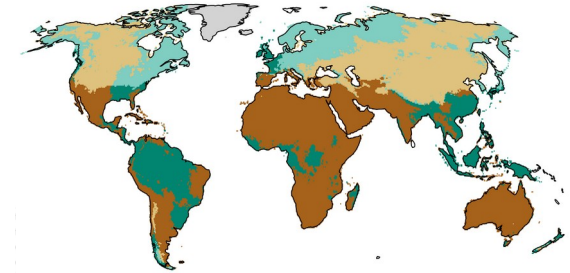
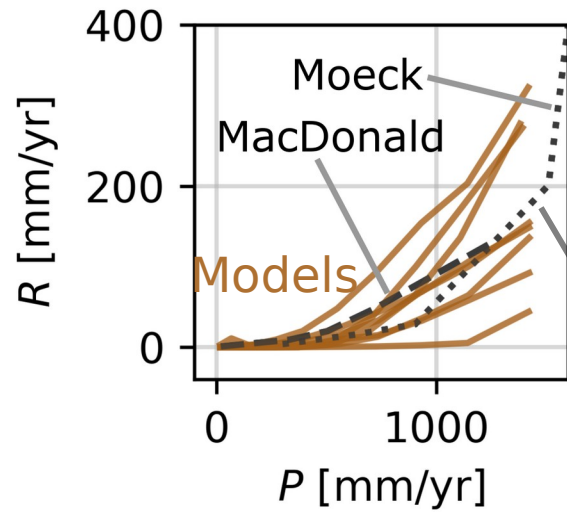
Examples of functional relationships



Examples of functional relationships

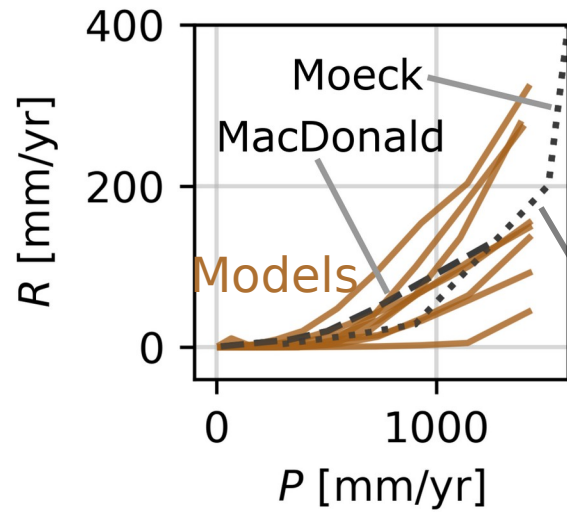


Precipitation-recharge relationships differ widely between models and observations

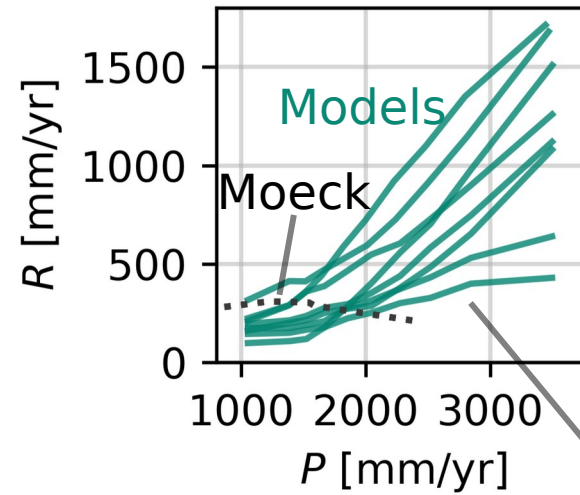


Models mostly
brace observed
relationship, but
differences are
huge

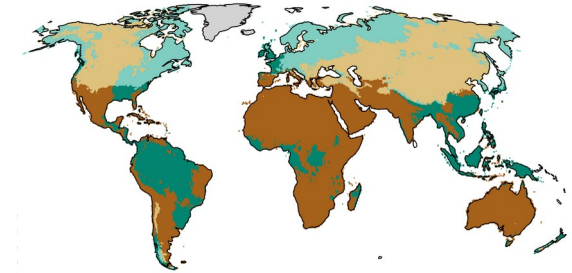
Precipitation-recharge relationships differ widely between models and observations



Models mostly
brace observed
relationship, but
differences are
huge



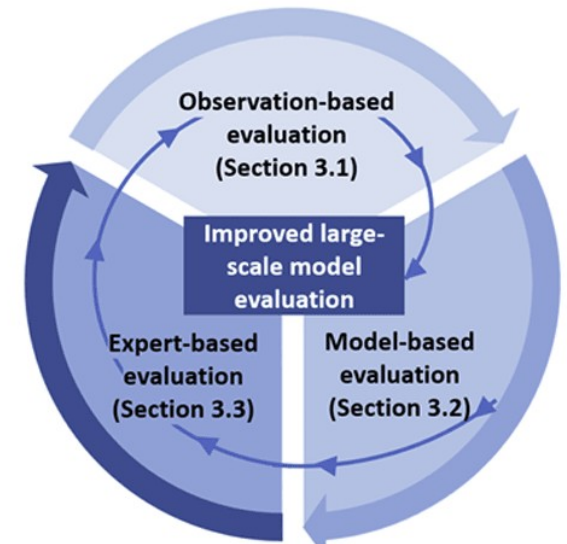
Models show
much larger
increase than
data suggest



There are many more evaluation approaches

- “Traditional” point comparison (e.g. using NSE)
- Elasticities (e.g. Zhang et al., 2023, Nature Water)
- Trends (e.g. Scanlon et al., 2018, PNAS)

→ We should think – as a community – about what each evaluation approach can and cannot do (e.g. Gleeson et al., 2021, GMD)



Community-driven model evaluation frameworks

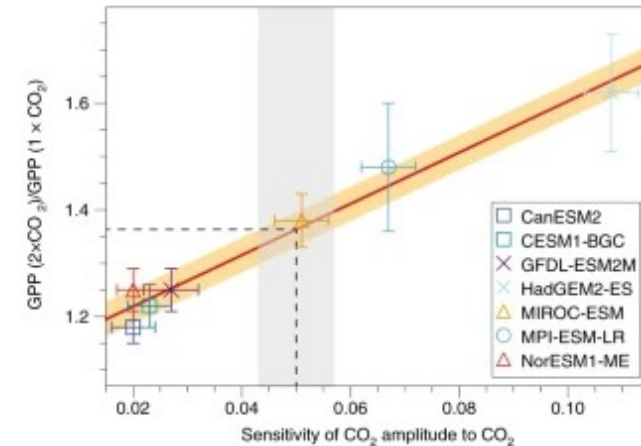
PERSPECTIVE

<https://doi.org/10.1038/s41558-018-0355-y>

nature
climate change

Taking climate model evaluation to the next level

Veronika Eyring ^{1,2*}, Peter M. Cox ³, Gregory M. Flato⁴, Peter J. Gleckler⁵, Gab Abramowitz ⁶, Peter Caldwell⁵, William D. Collins ^{7,8}, Bettina K. Gier^{1,2}, Alex D. Hall⁹, Forrest M. Hoffman ^{10,11}, George C. Hurtt¹², Alexandra Jahn ¹³, Chris D. Jones ¹⁴, Stephen A. Klein ⁵, John P. Krasting¹⁵, Lester Kwiatkowski ¹⁶, Ruth Lorenz ¹⁷, Eric Maloney ¹⁸, Gerald A. Meehl¹⁹, Angeline G. Pendergrass ¹⁹, Robert Pincus ¹⁸, Alex C. Ruane²⁰, Joellen L. Russell ²¹, Benjamin M. Sanderson ¹⁹, Benjamin D. Santer⁵, Steven C. Sherwood ⁶, Isla R. Simpson¹⁹, Ronald J. Stouffer ²¹ and Mark S. Williamson³



Improve observational datasets

Identify process basis of emergent constraints

Model evaluation should lead to model improvement



Taking water model evaluation to the next level

Improved process understanding

Explore process basis of functional relationships and learn more about hydrology

Improved models

Figure out which models cannot capture relationships and why

Improved data sets

Learn where we have data gaps (in functional space) and thus should measure

Next steps towards a “theory of evaluation”

- Possible elements of an evaluation framework
 - Evaluation approaches and their process basis
 - How to use evaluation for model improvement
 - Benchmark evaluation datasets
- Discussion: benchmarks for ISIMIP models?
 - Variables (streamflow, recharge, evaporation)
 - Gradients (precipitation, radiation)
 - Automatic quality control



References

- Rockström, J., Gupta, J., Qin, D., Lade, S. J., Abrams, J. F., Andersen, L. S., ... & Zhang, X. (2023). Safe and just Earth system boundaries. *Nature*, 1-10.
- Gnann, S., Reinecke, R., Stein, L., Wada, Y., Thiery, W., Schmied, H. M., ... & Wagener, T. (2022). Functional relationships reveal differences in the water cycle representation of global water models. *Under Review*.
- Gleeson, T., Wagener, T., Döll, P., Zipper, S. C., West, C., Wada, Y., ... & Bierkens, M. F. (2021). GMD Perspective: The quest to improve the evaluation of groundwater representation in continental to global scale models. *Geoscientific Model Development Discussions, 2021*, 1-59.
- MacDonald, A. M., Lark, R. M., Taylor, R. G., Abiye, T., Fallas, H. C., Favreau, G., ... & West, C. (2021). Mapping groundwater recharge in Africa from ground observations and implications for water security. *Environmental Research Letters, 16*(3), 034012.
- Wagener, T., Reinecke, R., & Pianosi, F. (2022). On the evaluation of climate change impact models. *Wiley Interdisciplinary Reviews: Climate Change, 13*(3), e772.
- Zhang, Y., Zheng, H., Zhang, X., Leung, L. R., Liu, C., Zheng, C., ... & Blöschl, G. (2023). Future global streamflow declines are probably more severe than previously estimated. *Nature Water, 1*(3), 261-271.
- Scanlon, B. R., Zhang, Z., Save, H., Sun, A. Y., Müller Schmied, H., Van Beek, L. P., ... & Bierkens, M. F. (2018). Global models underestimate large decadal declining and rising water storage trends relative to GRACE satellite data. *Proceedings of the National Academy of Sciences, 115*(6), E1080-E1089.
- Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., ... & Williamson, M. S. (2019). Taking climate model evaluation to the next level. *Nature Climate Change, 9*(2), 102-110.
- Eker, S., Rovenskaya, E., Obersteiner, M., & Langan, S. (2018). Practice and perspectives in the validation of resource management models. *Nature Communications, 9*(1), 5359.

