Title: Machine learning to generate gridded extreme precipitation data sets for global land areas with limited in situ measurements

Authors and Affiliations: Mark Risser (LBNL), Alan Rhoades (LBNL), Ankur Mahesh (UCB)

Focal Area: Data acquisition and assimilation enabled by machine learning, AI, and advanced methods including experimental/network design/optimization, supervised learning (including deep learning), and hardware-related efforts involving AI (e.g., edge computing).

Science Challenge | *Short statement describing the area addressed by the white paper.*

There is a strong need for gridded observational data sets that describe the climatology of extreme precipitation across the globe; such data sets are critical for quantifying precipitation extremes and their corresponding influence on large perturbations on surface and groundwater systems and flooding. Machine learning (ML) methods can be used to generate pseudo in situ measurements of the climatology of extreme precipitation for land regions with poor sampling by matching relevant geographic (e.g., orography) and atmospheric (e.g., surface temperature, precipitation, and pressure) variables for densely sampled regions and those with limited geographic sampling.

Rationale | *Description of the research needs/gaps, the barriers to progress, and the justification for and benefits associated with the proposed approach.*

Recent methodological work led by Mark Risser (Risser et al., 2019a) developed a specialized statistical analysis to characterize the climatology of extreme precipitation over the contiguous United States (CONUS), a region with dense geographic sampling of daily precipitation, to generate a new gridded data product, with explicit uncertainty quantification, based on in situ measurements specifically tailored to extreme precipitation. Critically, the data set preserves the statistics of extreme precipitation observed at weather station locations, unlike some traditional daily gridded products which can underestimate extreme seasonal precipitation by as much as 30% (Pierce et al., 2021). The method and the corresponding data set have been used to characterize high-resolution trends over CONUS (Risser et al., 2019b), determine relationships between extreme precipitation and various modes of climate variability (Patricola et al., 2020; Risser et al. 2021), validate gridded data products based on radar measurements (Molter et al., 2021, in prep.), and derive observationally-based attribution statements for changes in extreme precipitation (Risser et al. 2021b, in prep). However, such a data set can only be developed for land regions that are well-sampled geographically with (in the case of evaluating trends) up to century-length records. This criterion rules out much of the globe, including high-latitude regions of North America.

While daily precipitation is notoriously heterogenous, the statistics of extreme precipitation are much more well-behaved and well-understood. The punchline of this work will be to use machine learning to translate information on the climatology of extreme precipitation from a well-sampled region, like CONUS, to poorly-sampled regions, like Alaska. The most ideal solution would be to instead implement a new network of monitoring stations in these regions; however, a comprehensive space-time network of in situ measurements is extremely expensive and would take many years to yield enough information to characterize trends. Instead, or at least in the meantime, AI and ML methods can serve as a substitute to new measurements of the Earth system.

DOI: https://doi.org/10.2172/1769784
DOE BER, AI4ESP 2021 #AI4ESP1113 (https://www.ai4esp.org)

In spite of the fact that our focus is on the climatology of extreme precipitation (as opposed to individual events or year-to-year variability), extrapolating information from a well-sampled land region to a new unobserved region is not straightforward. The relationships between extreme precipitation and the many factors that drive extreme precipitation (e.g., orography, storm systems, and responses to large-scale climate modes of variability) are complex, and in many cases (especially with respect to extremes) cannot be derived from Global Circulation Models (GCMs) with horizontal resolutions that are too coarse to capture the relevant phenomena. As such, ML methods are essential to reveal and derive these relationships such that the climatology of extreme precipitation can be appropriately extrapolated.

Narrative | *Scientific and technical description of the opportunities and approach; activities that will advance the science; and specific field, laboratory, model, synthesis, and/or analysis examples*.

In order to generate pseudo-observations of the climatology of extreme precipitation, we will use ML methods on data sources (e.g., reanalysis and topographic variables) that exist for both CONUS and a new region of interest; our target will be to develop extreme precipitation data sets for Alaska. Alaska is chosen for several reasons: Alaska is (1) a major wildlife refuge and hence faces important vulnerabilities due to climate change; (2) facing an accelerated version of climate change relative to other parts of North America (Duarte et al., 2012); and (3) will emerge as a major shipping route as sea ice extent diminishes (Wei et al., 2020). Furthermore, there are important relationships between extreme precipitation, snowpack, and permafrost linkage, for example related to methane emission feedbacks (Douglas et al., 2020). However, there are serious limitations with the available in situ measurements in Alaska: first, the GHCN network only has 986 stations (CONUS, by contrast, has 21,308), in spite of the fact that Alaska is approximately 20% the size of CONUS (8,080,464 km^2 for CONUS and 1,717,856 km^2 for Alaska). Furthermore, only about 30 of the Alaska stations have century-length records with at least 66% non-missing daily data (compared to more than 2000 for CONUS; see Figure 1). Finally, the Alaska stations are not representative of the state's climatological zones: of the 986 stations, 98% are below 1000 meters above sea level (the highest elevations in Alaska significantly exceed 1500m).

Using CONUS and existing Alaskan measurements as training data, a neural network can be trained to increase the geographic coverage of a climatological gridded extreme precipitation data set. The input to our neural network will be topographic variables and reanalysis data, and the target for the network will be the Risser et al. (2019a) gridded climatology of extreme precipitation. Our model is trained to learn the relationship between the input and target on a spatial and temporal subset of the CONUS data, and then validated on the remainder of CONUS data and the entirety of data available from the stations in Alaska. As the input and target are both spatial datasets (with a latitude and longitude dimension), we propose using the U-Net neural network architecture (Ronneberger et al., 2015), as it is designed to handle inputs and targets of this spatial nature. Through numerical optimization of its parameters, the U-Net architecture extracts a compact representation of the input, and then it learns a mapping between this representation and the target. We aim to compare the U-Net architecture to deeper, more memory-intensive architectures, such as DeepLabv3+ (Chen et. al., 2018). This architecture is explicitly designed to learn information at multiple spatial scales, so it may be ideal for learning from station-level observations, gridded reanalysis datasets, and gridded topological data. DeepLabv3+ has been shown to be an effective architecture for learning from climate science datasets, as it has been used to identify atmospheric rivers and tropical cyclones (Kashinath et. al., 2021).

In summary, our main goal is to use information we have from the CONUS (a heavily instrumented region) to apply to an out of sample region like Alaska, since a comprehensive network of observations in space and time is a highly expensive task. One benefit of our methodology is that the accompanying measures of spatial uncertainty can inform where potential meteorological citing locations, or ARM campaign deployments, might help minimize estimate uncertainties. An ultimate test of our method will be to compare our estimates with high-frequency measurements from the DOE Office of Science Atmospheric Radiation Measurement (ARM) user facility, specifically the fixed observation site in the north slope of Alaska.

Over the next decade, we aim to tackle two fundamental problems in using machine learning to create an extreme precipitation dataset: (1) distribution shift and (2) continuously streaming data. First, anthropogenic climate change has influenced extreme precipitation in the continental United States (Kirchmeier-Young, 2021). As the return period of extreme precipitation events changes, we aim to account for this distribution shift as we train the neural network. For instance, we aim to leverage methods that allow the neural network to learn a residual function (Long et. al., 2016) between historical and present precipitation, in order to preserve the neural network's accuracy in the presence of climate change. Second, over the next decade, new stations will come online, and existing stations will collect more data. We plan to incorporate the continuous stream of new observations to train our neural networks, particularly because newer data offers crucial training samples that characterize the aforementioned distribution shift. It is computationally infeasible to retrain the neural network every time a new observation becomes available, particularly at daily time scales. Online learning methods (Hoens et. al., 2012) allow for neural networks to dynamically update their weights using a window of the most recent predictions and a learned weighting scheme between older and newer data points.

Regarding FAIR (Findable, Accessible, Interoperable, Reusable) principles, as with related work (e.g., Risser et al., 2019a) all software for the specialized spatial extreme value analysis is made freely available, with reproducibility of results being a core element of the research conducted. Similarly, all resulting data products (e.g., Risser et al., 2019c) will be made publicly available through robust data-sharing platforms. Finally, as our various analyses utilize the Python (https://www.python.org/) and R (https://cran.r-project.org/) programming languages, all code is open source.

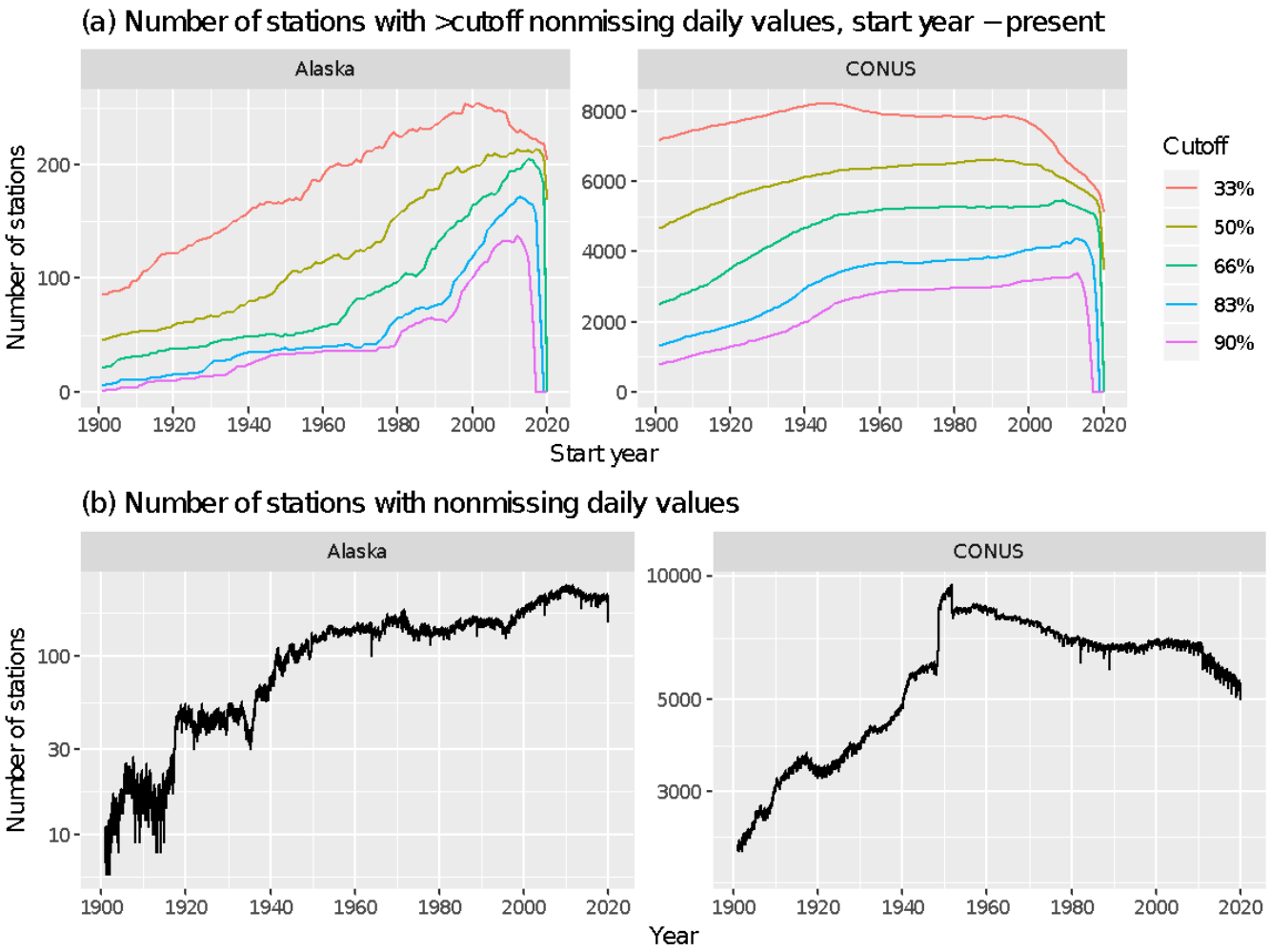

Figure 1: Comparison of the number of GHCN gauge measurements of daily precipitation for Alaska and CONUS. Top: Number of stations with at least a certain percentage of non-missing daily values for a range of start years to present. Bottom: For each day, the number of stations with non-missing daily values.

References (Optional)

Risser, M.D., Paciorek, C.J., O'Brien, T.A., Wehner, M.F., Collins, W.D. (2019a). A probabilistic gridded product for daily precipitation extremes over the United States. *Climate Dynamics*, 53(5):2517--2538. DOI: 10.1007/s00382-019-04636-0

Pierce, D.W., Su, L., Cayan, D.R., Risser, M.D., Livneh, B., Lettenmaier, D.P. (2021). An extreme-preserving long-term gridded daily precipitation data set for the conterminous United States. *Under review*.

Risser, M.D., Paciorek, C.J., O'Brien, T.A., Wehner, M.F., Collins, W.D. (2019b). Detected changes in precipitation extremes at their native scales derived from in situ measurements. *Journal of Climate*, 32(23), 8087-8109. DOI: 10.1175/JCLI-D-19-0077.1

Risser, M.D., Paciorek, C.J., O'Brien, T.A., Wehner, M.F., Collins, W.D. (2019c). A probabilistic gridded product for daily precipitation extremes over the United States, https://doi.org/10.7910/DVN/LULNUQ, Harvard Dataverse, V3.

Risser, M.D., M. Wehner, J. P. O'Brien, C. Patricola, T. O'Brien, W. Collins, C. Paciorek, H. Huang. (2021). Quantifying the influence of natural climate variability on in situ measurements of

seasonal total and extreme daily precipitation. *Climate Dynamics*. DOI: 10.1007/s00382-021-05638-7

Patricola, C.M., O'Brien, J.P., Risser, M.D., Rhoades, A.M., O'Brien, T.A, Ullrich, P.A., Stone, D.A., Collins, W.D. (2020). Maximizing ENSO as a source of western US hydroclimate predictability. *Climate Dynamics*, 54(1-2), 351-372. DOI: 10.1007/s00382-019-05004-8

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." *International Conference on Medical image computing and computer-assisted intervention*. Springer, Cham, 2015.

Kirchmeier-Young, Megan C., and Xuebin Zhang. "Human influence has intensified extreme precipitation in North America." Proceedings of the National Academy of Sciences 117.24 (2020): 13308-13313.

Kashinath, Karthik, et al. "ClimateNet: an expert-labeled open dataset and deep learning architecture for enabling high-precision analyses of extreme weather." Geoscientific Model Development 14.1 (2021): 107-124.

Duarte, C. M., Lenton, T. M., Wadhams, P., & Wassmann, P. (2012). Abrupt climate change in the Arctic. Nature Climate Change, 2(2), 60-62.

Wei, T., Yan, Q., Qi, W., Ding, M., & Wang, C. (2020). Projections of Arctic sea ice conditions and shipping routes in the twenty-first century using CMIP6 forcing scenarios. Environmental Research Letters, 15(10), 104079.

Douglas, T. A., Turetsky, M. R., & Koven, C. D. (2020). Increased rainfall stimulates permafrost thaw across a variety of Interior Alaskan boreal ecosystems. NPJ Climate and Atmospheric Science, 3(1), 1-7.

Chen, Liang-Chieh, et al. "Encoder-decoder with atrous separable convolution for semantic image segmentation." Proceedings of the European conference on computer vision (ECCV). 2018

Hoens, T. Ryan, Robi Polikar, and Nitesh V. Chawla. "Learning from streaming data with concept drift and imbalance: an overview." Progress in Artificial Intelligence 1.1 (2012): 89-101.

Long, Mingsheng, et al. "Unsupervised domain adaptation with residual transfer networks." *arXiv preprint arXiv:1602.04433* (2016).