scientific reports

OPEN

Evaluating the ability of gridded climate datasets to capture temperature and precipitation trends and extremes

Erik Smith[3✉], Nina Grant[1,3], Xinyue Luo[2,3] & Delavane Diaz[3]

Data informs policy and drives decisions. With an increasing number of regulations requiring entities to use local climate data in their planning, it's more important than ever to understand the strengths and limitations of data we use. While they have been shown to capture long-term statistics on global or regional levels, the ability of gridded climate datasets to capture trends and extreme events is not common knowledge. Four widely used gridded datasets, ERA5, ERA5-Land, MERRA-2, and PRISM, were assessed for their ability to capture extreme heat, extreme cold, and heavy precipitation events, as well as trends in annual maximum and minimum temperatures and total precipitation, over the contiguous US (CONUS). Spatial patterns are evident in each dataset, with the largest differences between observations and the gridded data across the western US for temperature and along the Gulf Coast for heavy precipitation. In general, gridded datasets captured extreme heat better than extreme cold or heavy precipitation, and trends in annual maximum temperature better than trends in annual minimum temperatures and annual total precipitation. All dataset capture extreme heat days comparably, but PRISM generally performed best for extreme cold and the bias-adjusted MERRA-2 dataset generally performed best for heavy precipitation days.

Gridded climate datasets, which includes reanalysis products and statistically derived products, are used in numerous applications including detection and attribution studies, climate model validation, power system planning, renewable energy analysis, agricultural modeling, and more. As of Jan 2024, the NCEP-NCAR dataset, the first reanalysis dataset created by NOAA, had been cited more than 34,000 times[1,2]. Despite their prolific usage, gridded datasets are not actual observations and the process by which gridded datasets are created makes errors inherent. However, gridded datasets are not meant to replace surface observations, rather they are meant to fill spatial and temporal gaps between observations. So, while differences between surface observations and gridded data can be misinterpreted as errors, they are simply different and should be interpreted as so. Nonetheless, it is important to quantify these differences so that end-users of gridded data products can better understand appropriate applications as well as the uncertainties that may arise from the use of gridded datasets.

Uncertainty in gridded datasets can arise from several sources including quality of the observational data, the density or structure of the observational network, and the interpolation or assimilation method chosen[3–5]. Structural uncertainty (data selection and analysis method) has been found to have the greatest effect on extremes in gridded datasets[3,6]. Efforts to evaluate gridded datasets use techniques and metrics such as cross-validation, ensemble uncertainty estimates, probability distribution functions, and spatial correlation[2,7–9]. However, most studies focus on the skill at capturing the means, not the extremes, and focus on annual, seasonal, or monthly scales as opposed to daily or sub-daily time scales[8].

Past assessments of gridded dataset skill in capturing extremes have found that by looking only at global spatial scales or annual to monthly timescales can obscure major differences between datasets[4,7,8]. While most gridded datasets are consistent for mean temperatures, other moments like variance and skewness are not robust across datasets[7]. These studies have also found that that method of calculation (interpolation or assimilation method, observational network used, observational record length, order of operation, etc.), especially the underlying observational network, has the largest effect on results, while changes to specific parameters have little effect[3–5]. Most studies concluded that it is best to use different datasets for different research questions (such as trend analysis, instantaneous field estimates, regional studies, particular variables of interest), and, where the network is sparce or datasets inconsistent, it is recommended that a range be determined rather than averaging

[1]Rutgers University, New Brunswick, USA. [2]Rice University, Houston, USA. [3]Electric Power Research Institute, Palo Alto, USA. ✉email: esmith@epri.com

the reanalyses[3–5,8–12]. However, using multiple climate datasets is often unfeasible, as these datasets are quite large and require considerable computing power to process. Thus, knowing the most applicable dataset for the purpose is beneficial.

Each gridded climate dataset goes through rigorous validation, however, these validations are typically not done with the intent of providing an end-user the necessary information for selecting the optimal dataset for their application. Moreover, there is little peer-reviewed guidance for those using gridded datasets on how to evaluate the available datasets and which datasets may be best for specific regions or for specific climate hazards. With more companies conducting physical risk assessments, knowing which datasets are available and how well they represent extremes and trends in the area of interest is critical. In this paper, we assess the ability of four gridded datasets, ERA5, ERA5-Land, MERRA-2, and PRISM, to capture the magnitude and timing of extreme heat, extreme cold, and heavy precipitation, as well as trends in annual maximum and minimum temperatures and annaual total precipitation, over the contiguous United States (CONUS). The magnitude of the differences are compared across datasets for each surface station to show dataset bias by location and hazard. This analysis provides insights on the differences between gridded datasets and surface observations when it comes to representing extreme weather events at the local level. These results can help determine which dataset may be best to use if assessing extreme temperature or heavy precipitation events.

Methods
Weather data
Observational weather station data were retrieved using the meteostat Python library (https://dev.meteostat.net/python/), which accesses publicly available surface station data from the National Oceanic and Atmospheric Administration. Only US stations with a temporal coverage of at least 80% from 1981 to 2022 were included. This resulted in 317 stations across the contiguous US (CONUS) for temperature and 267 stations for precipitation.

The gridded datasets chosen for this study include ERA5[13], ERA5-Land[14], PRISM[15], and MERRA-2[16]. ERA5 was developed by the ECMWF Copernicus Climate Change Service as the successor to ERA-Interim[13]. ERA5 has the longest temporal coverage of the datasets, with hourly data extending back to 1940 and a spatial resolution of 31 km. ERA5-Land is a 9 km land-only model of ERA5, forced with the ERA5 atmospheric output[14]. MERRA-2 is the second iteration of MERRA global reanalysis dataset. Unlike the other datasets, MERRA-2 includes bias-adjusted precipitation data, but has the lowest resolution of the selected datasets at approximately 50 km[16]. PRISM is an interpolated gridded dataset calculated with a climate-elevation regression based on location, elevation, topography, orography, proximity to coasts, the vertical atmospheric layer, and includes 10,000 spatially quality-controlled temperature surface stations and 13,000 precipitation stations[15]. It has the highest resolution of our selected datasets (4 km) and is the only dataset based on statistical interpolation of surface station observations[15]. Conversely, reanalysis data is created by assimilating various types of observational data into numerical weather models[1]. In other words, reanalysis data relies on physical principals to create spatially and temporally continuous datasets, whereas interpolated data relies on statistical methods.

The grid points from the gridded datasets overlapping the surface stations were used to compare events. Land-only datasets (PRISM and ERA5-Land) did not always overlap certain coastal stations. Thus, the neighboring grid cells were assessed for the best match to the surface station. Island and buoy locations were removed from the station list.

Metrics
Metrics were created based on their relevance to industry. In other words, these metrics are commonly used and well understood by end-users, thus they are an important point of comparison[17]. Extreme heat days (EHDs) are defined as daily maximum temperatures at or above the summer (June – August, JJA) 95th percentile and extreme cold days (ECDs) as daily minimum temperatures at or below the winter (December-February, DJF) 5th percentile. Heavy precipitation days (HPD) are defined as days in which the daily total precipitation exceeds the 99th percentile (all days). Percentiles are calculated using the station data and based on the current climate normal period, 1991–2020.

Mean absolute error (MAE) is used as the primary metric to quantify the ability of each gridded product to capture the magnitude of the extreme event at the surface stations. The MAE is calculated between the station observations and gridded data for only the days in which the station observation meets or exceeds ECD (5th percentile), EHD (95th percentile), or HPD (99th percentile) threshold. For precipitation, mean absolute percentage error (MAPE) is used to account for the large range of daily precipitation values across the CONUS.

To assess timing of extreme temperature and heavy precipitation days, a match percentage is used. The match percentage of EHDs, ECDs, and HPDs is defined as the number of days in which the gridded datasets exceed the respective temperature or precipitation threshold divided by the number of days the station exceeded the threshold. Note, near misses are not allowed, thus locations with small annual temperature variations or low precipitation are likely to have a lower match percentage. False alarms, or days in which the gridded dataset exceeds the threshold when the surface station does not, are shown in the supplemental information. False alarms are shown using percentages, which are calculated by taking the number of days the gridded dataset shows an EHD, ECD, or HPD while the observations do not, and then dividing by the total number of EHD, ECD, and HPD from the observations respectively. For EHDs and ECDs, there are generally very few false alarms, however, HPDs have considerably more false alarm days.

Lastly, because trend estimation is critical for characterizing the changing risk to climate extremes, the ability of the gridded datasets to capture trends in each variable is assessed. A Theil-Sen slope estimation, which is robust to outliers[18], is used to estimate the decadal trend for the annual maximum temperatures, annual minimum temperatures, and annual total precipitation. The slope from the gridded datasets are then differenced

from the Theil-Sen slope estimate for the surface stations to assess differences in the magnitude and direction of the trends.

Results

Extreme heat days and annual maximum temperature trends

The EHD threshold at each surface station, or summer 95th percentile, is shown in Fig. 1A and the trends in annual maximum temperatures in Fig. 1B as a reference. All four gridded datasets capture EHDs comparably (Fig. 1). The MAE is generally lowest in the southeastern US, with larger differences across the Northeast and western US. MAEs also tend to be larger near coastlines, such as the Great Lakes and Atlantic Coast. MERRA2, which has the coarsest spatial resolution of all the datasets, tends to have lowest MAEs for EHDs, particularly across the central US and into the Rockies. Conversely, PRISM, which has the highest spatial resolution, tends to have higher MAEs, particularly in the north-central US.

While annual maximum temperatures have generally warmed across the US, the trend has been larger in the western US. Some stations across the Midwest and Southeast have even exhibited a cooling trend in annual maximum temperatures since 1981. Figure 1D shows the difference in.

decadal trends between the station and the gridded dataset. A positive difference generally means the gridded dataset had a larger positive trend since 1981 than the surface station. Conversely, a negative value means the trend in the gridded dataset was generally smaller or even opposite in sign than the observed trend. While all datasets capture the trends relatively well, with differences close to 0, PRISM was generally closest to the observed trend. ERA5, ERA5-Land, and MERRA2 all have a bias toward larger positive trends than was otherwise observed. For MERRA2, this is particularly true in the Northeast, while large trend differences in ERA5 and ERA5-Land are more scattered across CONUS.

Extreme cold days and annual minimum temperature trends

The ECD threshold at each surface station, or winter 5th percentile, is shown in Fig. 2A and the trends in annual minimum temperatures in Fig. 2B as a reference. As with EHDs, the MAE in ECDs is generally lowest in the southeastern US, with larger differences across the northern and western US (Fig. 2C). However, MAEs for ECDs tends to be larger than the MAEs for EHDs. MERRA2 generally has lower MAEs across the Southeast compared to ERA5 and ERA5-Land and PRISM, however, ERA5 and ERA5-Land tend to have lower MAEs across the northern US. MERRA2 has the highest MAEs across the western US, which may be due to the coarser spatial resolution of the dataset not capturing large variations in topographically diverse locations. PRISM, generally has the lowest MAEs in ECDs. The high resolution of PRISM appears to be particularly helpful in capturing local extremes in locations with complex topography, like along the Florida coast lines and across the mountain west. PRISM is also the only non-reanalysis dataset, thus the statistical methods by which the gridded data is created may anchor the data more closely to the surface observations and allow for a better match.

Compared to annual maximum temperatures, annual minimum temperatures have warmed considerably more across the US since 1981, with the largest positive trend across the eastern US. Figure 2D shows the difference in decadal trends between the station and the gridded dataset. A positive difference generally means the gridded dataset had a larger positive trend since 1981 than the surface station. Conversely, a negative value means the trend in the gridded dataset was generally smaller or even opposite in sign than the observed trend. The differences between the trend in gridded dataset and the observed trends are generally larger than for annual maximum temperature. This is particularly true across the Midwest in the ERA5, ERA5-Land,

and MERRA2 datasets, which overestimate the warming in annual minimum temperatures. Nonetheless, the gridded datasets generally capture the spatial distribution of the trends well. Again, PRISM most closely aligns with the observed trend, and while the largest differences in trends are also across the Midwest, there are fewer large differences when compared to the other three datasets.

Heavy precipitation days and annual total precipitation trends

The HPD threshold at each surface station, or 99th percentile, is shown in Fig. 3A and the trends in annual total precipitation in Fig. 3B as a reference. The MAPE for ERA5 and ERA5-Land is very similar across stations, with the lowest MAPEs across the Northeast and western US (Fig. 3C). MERRA2, which is a bias-adjusted precipitation dataset, has generally lower MAPEs than ERA5 and ERA5-Land across CONUS. So while MERRA2 is coarser than ERA5 and ERA5-Land, the bias-adjustment proves useful in capturing the magnitude of HPDs across the CONUS. However, it should be considered whether this improvement from bias-adjustment would hold for locations with no surface observations. Unlike EHDs and ECDs, PRISM tends to have the highest MAPE for HPDs across the CONUS. The statistical method PRISM uses to interpolate precipitation may be less capable of capturing the often highly localized nature of HPDs than reanalysis datasets, which are created with physical models.

Annual total precipitation has generally increased across the eastern US since 1981, with decreases predominant across the western US (Fig. 3B). As with annual maximum and minimum temperature trends, a positive difference generally means the gridded dataset had a larger positive trend since 1981 than the surface station while a negative value means the trend in the gridded dataset was generally smaller or even opposite in sign than the observed trend. For annual total precipitation, ERA5, ERA5-Land, and MERRA2 generally underestimated the decadal trend since 1981, particularly in the eastern US (Fig. 3D). PRISM, however, tended to overestimate the decadal trend in annual total precipitation since 1981. Though different units, gridded datasets generally don't capture trends in annual total precipitation as well as annual maximum and minimum temperature. In Lousiville, KY, for instance, the ERA5, ERA5-Land, and MERRA2 trends were off by approximately 40–45% (trends were + 37 mm, + 37 mm, and + 41 mm per decade respectively, as opposed to the observed + 67 mm).
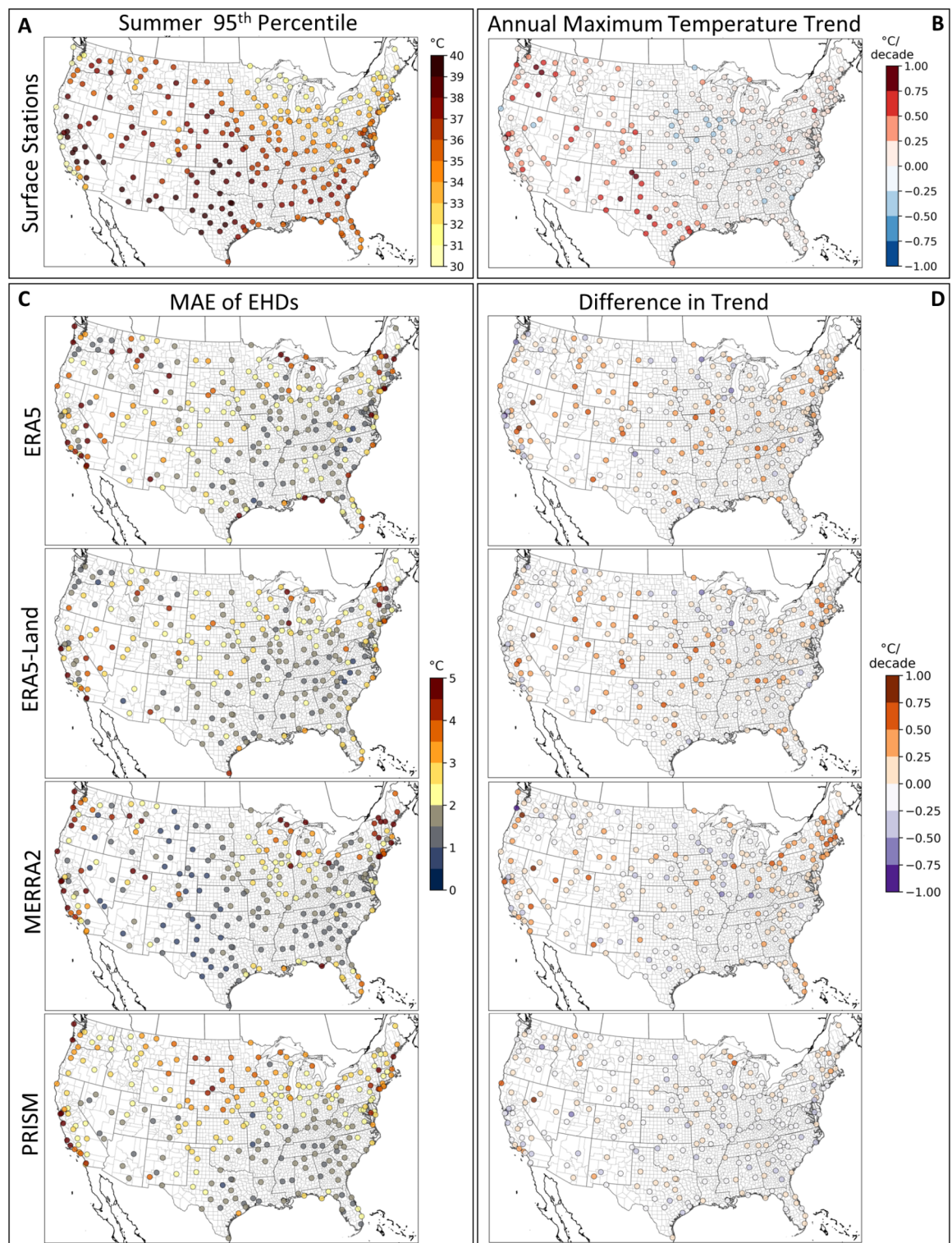
Fig. 1. (**A**) 95th percentile at each surface station, based on the 1991–2020 summer (JJA) daily maximum temperatures. (**B**) Decadal trend in annual maximum temperatures at each station. (**C**) Mean absolute error of extreme heat days (EHD; days > = 95th percentile) for each gridded dataset. (**D**) Difference in decadal trends in annual maximum temperatures for each gridded dataset (trend in gridded dataset minus observed trend). Figure created using python matplotlib.
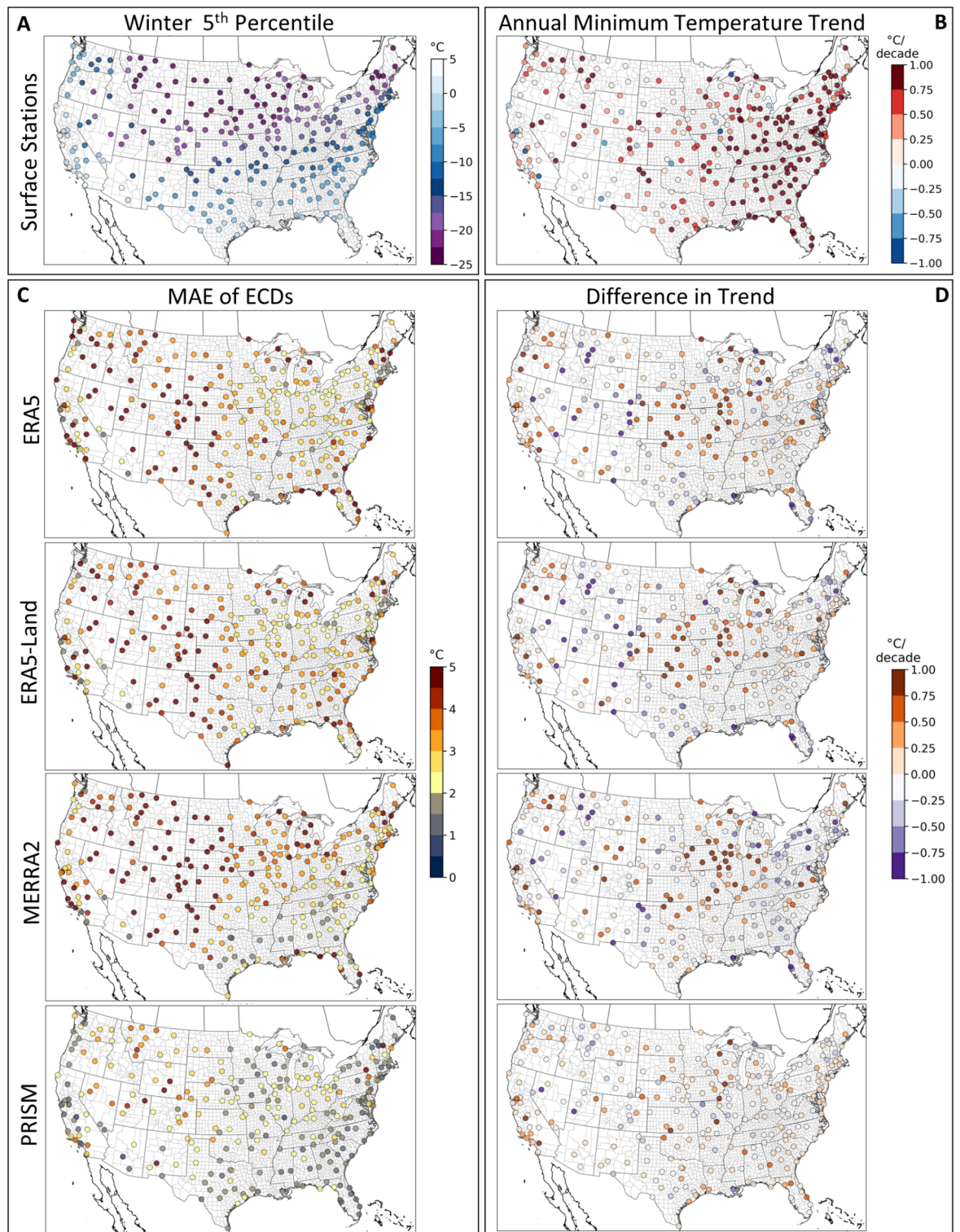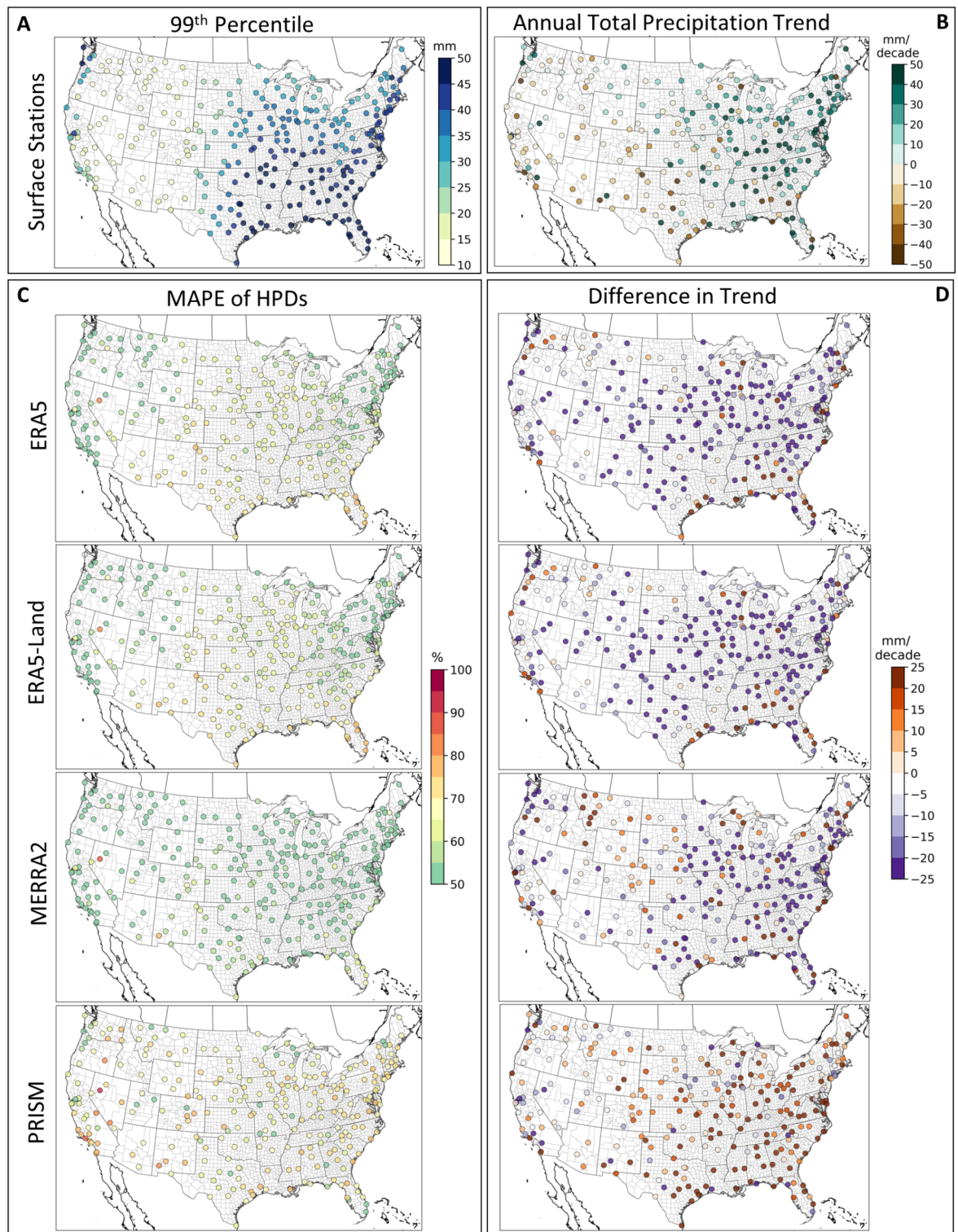
Fig. 2. (**A**) 5th percentile at each surface station, based on the 1991–2020 winter (DJF) daily minimum temperatures. (**B**) Decadal trend in annual minimum temperatures at each station. (**C**) Mean absolute error of extreme cold days (ECD; days < = 5th percentile) for each gridded dataset. (**D**) Difference in decadal trends in annual minimum temperatures for each gridded dataset (trend in gridded dataset minus observed trend). Figure created using python matplotlib.

Fig. 3. (**A**) Heavy precipitation day (99th percentile) threshold, based on the daily precipitation (mm) at each surface station over the 1991–2020 period. (**B**) Decadal trend in annual total precipitation at each station. (**C**) Mean absolute error of heavy precipitation days (HPD; days >= 99th percentile) for each gridded dataset. (**D**) Difference in decadal trends in annual total precipitation for each gridded dataset (trend in gridded dataset minus observed trend). Figure created using python matplotlib.

Comparison of datasets

Gridded datasets differ for a multitude of reasons. The physical (or statistical) model driving the data, the assimilation technique, input datasets (i.e., observational datasets, satellite and other measurements, etc.), parameterizations, output resolution of the data, and whether or not bias correction is applied to the data are all factors that contribute to differences between these datasets. These factors provide important context when comparing across datasets and can even provide insight into how individual gridded datasets may be improved. For example, bias-adjustment of the MERRA2 is shown to have considerable improvement in representing HPDs, thus post-processing of gridded datasets may present an opportunity for improvement. While these modeling questions are important considerations for data creators, an end-user will generally stop at the point of comparison. In other words, end-users need to know which available dataset is best-suited for the application. To this end, a comparison of datasets across metrics is of the utmost importance.

Combining the information into a single distribution, we can more easily visualize the biases across datasets and by variable (Fig. 4). The gridded datasets generally capture EHDs better than ECDs and HPDs, though PRISM tends to do a better capturing ECDs than EHDs. ERA5 and ERA5-Land have very similar distributions of MAE for EHDs and ECDs, and MAPE for HPDs, thus the increased spatial resolution for ERA5-Land doesn't appear to improve local representation of EHDs, ECDs, or HPDs. The The 95th percentile MAE for ECDs for MERRA2 is considerably higher than the other datasets, which indicates the coarse resolution may be a limitation in mountainous locations, where cold temperatures along narrow mountain ridges or cold sinks in mountain valleys may not be captured. While PRISM captures EHDs comparable and ECDs better than the other three datasets, it has the higher MAPE for precipitation. Though PRISM has the highest spatial resolution of the gridded datasets, the statistical methods being used do not appear to perform as well as physical models for capturing heavy precipitation. Moreover, the bias-adjusted MERRA2 performs best for HPDs though it has the coarsest spatial resolution.

Trends in annual maximum and minimum temperatures are generally captured better than trends in annual total precipitation. PRISM has a tighter distribution for annual maximum and minimum temperature trends, which means the trends are closer to observed for more locations. Trends in annual minimum temperatures are generally more different from the observed trends compared to trends in annual maximum temperatures.
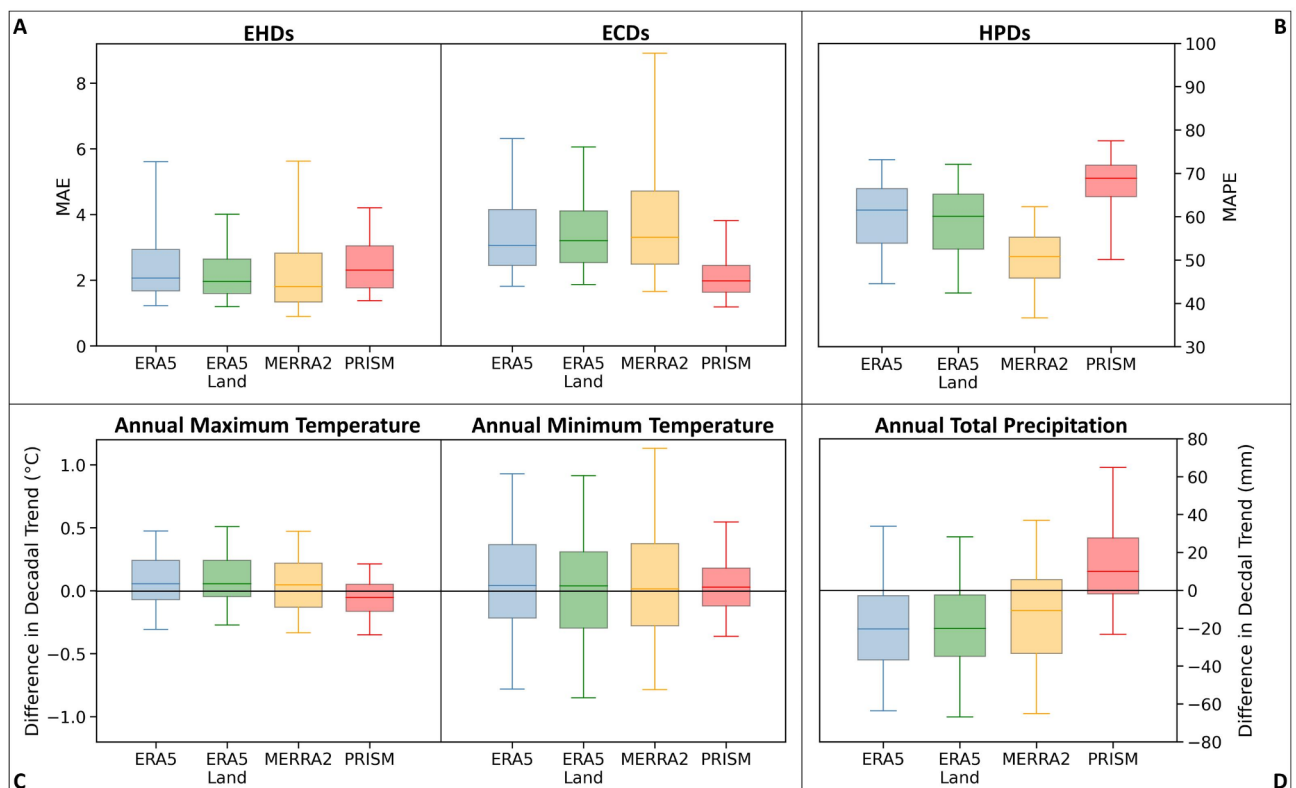

Fig. 4. (**A**) Boxplot showing the distribution of MAE for extreme heat days (EHDs) and extreme cold days (ECDs) for all points and each gridded dataset. (**B**) Boxplot showing the distribution of MAPE for heavy precipitation days (HPDs) for all points and each gridded dataset. (**C**) Boxplot showing the distribution of differences in decadal trends in annual maximum and minimum temperature for all points and each gridded dataset. (**D**) Boxplot showing the distribution of differences in decadal trends in annual total precipitation for all points and each gridded dataset. The whiskers of the boxplots show the 5th (lower whisker) and 95th percentile (upper whisker). Outliers are not shown to put the focus on central tendencies. Figure created using python matplotlib.
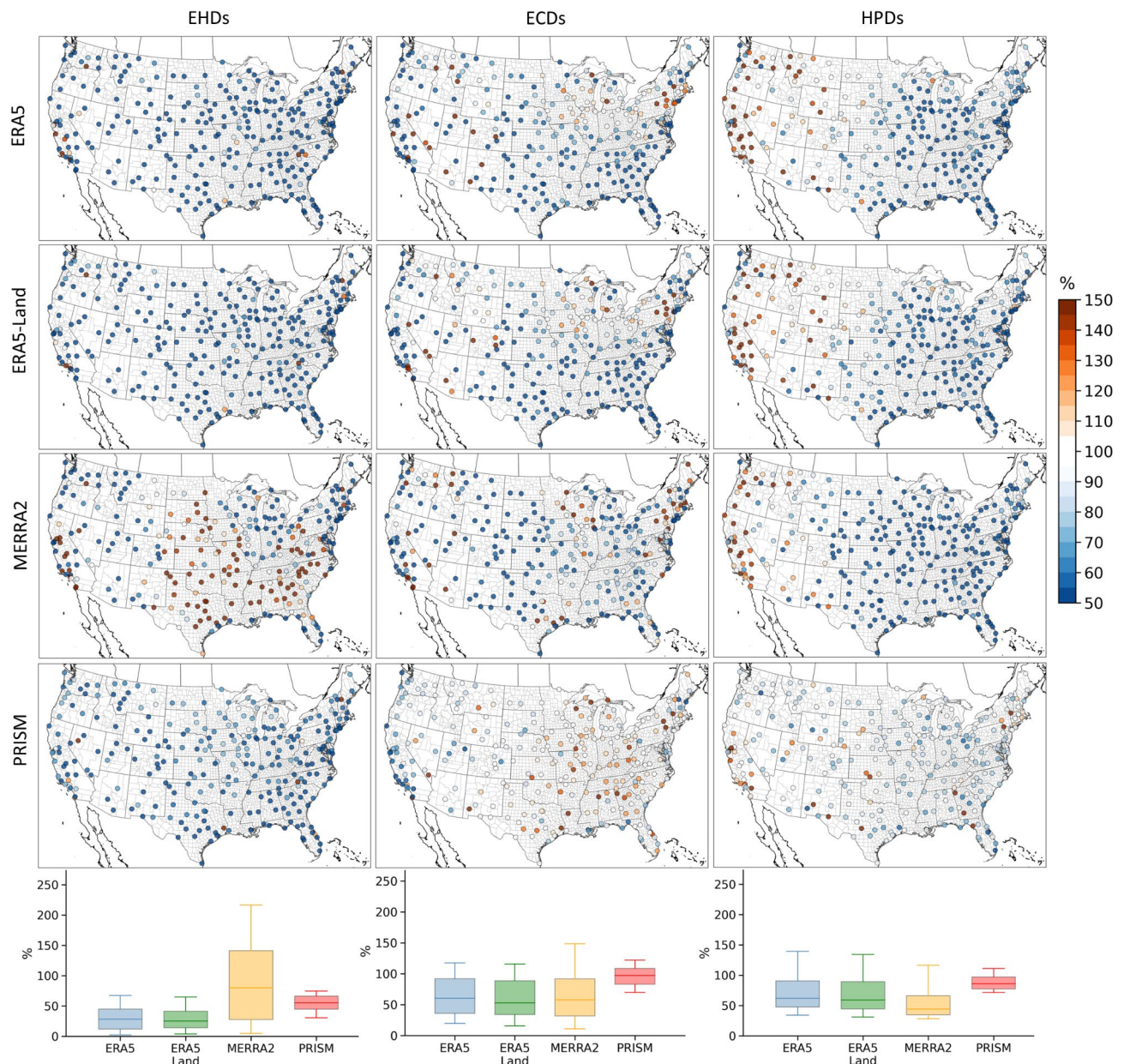
Fig. 5. Top - match percnetages by station for extreme heat days (EHDs), extreme cold days (ECDs), and heavy precipitation days (HPDs). Bottom - boxplot showing the distribution of match percnetages for extreme heat days (EHDs), extreme cold days (ECDs), and heavy precipitation days (HPDs) for all points and each gridded dataset. Figure created using python matplotlib.

This may be explained in part because the trends in annual minimum temperatures are generally larger thant the trends in annual mximum temperatures, thus more room for error. Nonetheless, gridded datasets typically capture trends in annual minimum and maximum temperaures well. While PRISM has the highest MAPE for HPDs, it tends to capture trends in annual total precipitation better than the other three gridded datasets. Both MERRA2 and PRISM capture the trends better than ERA5 and ERA5-Land. PRISM tends to overestimate the trend in annual total precipitation since 1981 while MERRA2, ERA5, and ERA5-Land tend to underestimate the trends.

In contrast to[19], which focused on apparent temperature in gridded datasets, we find the largest differences in temperature-only calculations for ECDs. The inclusion of humidity and 10-m wind speeds in the apparent temperature calculation may contribute to the differences in the findings. Other studies have also found larger discrepancies in the cold extremes[7,8,20]. Like previous studies[7], this research suggests additional studies are needed to determine the reason cold extremes generally exhibit larger differences from surface observations.

While it is important that gridded climate datasets capture the magnitude of extreme temperature and heavy precipitation events, it's also important to know if they are able to match the frequency of these events. Match percentages are used to assess the ability of gridded datasets to get the timing right on these events (Fig. 5).

Gridded datasets generally capture the timing of ECDs better than EHDs. In other words, more ECDs are captured by the gridded datasets than the EHDs. One reason for this may be that cold air is more dense and potentially more uniform across large extents, thus the coarseness of the gridded dataset is better able to capture these events compared to EHDs. For some locations, MERRA2 tends to produce more EHDs than actually occurred (match % > 100). All datasets have several locations with a match percentage above 100% for ECDs, but not nearly as many days as MERRA2 does for EHDs. PRISM generally has the highest match percentage for ECDs and HPDs. ERA5, ERA5-Land, and MERRA2 have lower match percentages than PRISM for ECD and HPDs, particularly in the western US for ECDs and in the east for HPDs. The lower match percentage for HPDs in the eastern US (for ERA5, ERA5-Land, and MERRA2) is likely due to these locations having a higher HPD threshold than drier locations generally found in the western US.

Conclusion

Surface observations are the most accurate records we have of historical weather, but the spatial gaps, missing data, and inconsistent lengths of record create challenges for numerous applications. Gridded datasets were created to address these challenges, but they are inherently different from point observations. Most notably, gridded datasets are meant to represent a much larger area than point observations. In other words, a 55 km grid cell is tasked with representing a much larger and potentially more diverse area than a single measurement from a surface station. Thus, when gridded datasets differ from surface stations, these differences should not be viewed as errors. Conversely, surface observations are a key input into gridded datasets and these observations will intrinsically weight the value of the grid cell. For this reason, gridded data is often a very close match to surface station observations.

The results of this study suggest that while all four gridded datasets do reasonably well in reproducing extreme events, some datasets are better able to reproduce extremes in certain regions and for specific hazards. All four datasets reproduce the magnitude and timing (match percentage) of extreme heat days (EHDs) comparably, but PRISM most closely reproduces the magnitude and timing of extreme cold days (ECDs). The gridded datasets also tend to reproduce extreme temperature events in the eastern US better than the western US. Complex terrain and microclimate in the western US create challenging dynamics that are difficult to capture in coarse gridded datasets. The bias-adjusted MERRA2 precipitation data captures heavy precipitation days (HPDs) better than ERA5 and ERA5-Land across the US, with PRISM generally having the largest MAPEs. This suggests bias-adjustment can overcome at least some of the limitations with representing precipitation in coarse datasets. PRISM's larger differences may be attributed to the statistical interpolation being used as opposed to a physical model.

Though not common knowledge, using gridded datasets for climate trends is generally not recommended. However, this research shows that while the trends in annual maximum temperatures are different, they are generally very close to the observed trends from stations. The trends in annual minimum temperatures, however, have larger differences and a consistent positive bias, particularly in the Midwest. That is, the trends in the gridded datasets tend to show more warming than was otherwise observed at the surface station, thus one should use caution when estimating trends in extreme cold from gridded data. PRISM tends to have the closest match to observed trends in annual minimum temperatures. There are larger biases in the trends in annual total precipitation from the gridded datasets than for annual maximum and minimum temperature. This is to be expected as precipitation is less continuous than temperature and driven by small-scale features, like topography, and complex processes, like convection. ERA5, ERA5-Land, and MERRA2 all tend to have smaller trends than was observed east of the Rockies, while PRISM tends to have larger trends than was observed.

While reanalysis products have been shown to represent averages well[7,8], these results suggest that caution should be used when using gridded datasets to characterize extremes. More specifically, extreme cold and heavy precipitation events may not be as well captured at the local level in gridded datasets as extreme heat events. For a more robust analysis of extremes, it would be best to quantify the range of uncertainty across multiple gridded datasets as well as nearby surface observations. It would also be beneficial to compare gridded datasets to indepdent observational sources. Future work will add additional gridded datasets, such as the recently released CONUS404 data, which is a downscaled ERA5 dataset aimed at hydrological applications[21]. Future work will also explore the use of machine learning for improving gridded datasets.

Data availability

The data used in this study can be found online for download at the following locations. ERA5 and ERA5-Land daily surface temperature and precipitation data are available from the European Center for Medium-Range Weather Forecasts (ECMWF) via the Copernicus Climate Change Service (at and https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=form). PRISM air temperature and precipitation data are available from the PRISM Climate Group at Oregon State (at https://prism.oregonstate.edu/downloads/). MERRA-2 data is made available on the National Aeronautics and Space Administration (NASA) Goddard Earth Sciences Data and Information Services Center (GES DISC). The code developed for this study can be made available upon request. The datasets used and/or analysed during the current study will be made available from the corresponding author on reasonable request.

References

1. Kalnay, E. et al. The NCEP/NCAR 40-Year reanalysis project. *Bull. Am. Meteorol. Soc.* **77**(3), 437–472. DOI: 10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2, (1996).
2. Parker, W. S. Reanalyses and observations: What's the difference?? *Bull. Am. Meteorol. Soc.* **97**(9), 1565–1572. https://doi.org/10.1175/BAMS-D-14-00226.1 (2016).
3. Dunn, R. J. H., Donat, M. G. & Alexander, L. V. Investigating uncertainties in global gridded datasets of climate extremes. *Clim. Past* **10**(6), 2171–2199. https://doi.org/10.5194/cp-10-2171-2014 (2014).
4. Ge, S., Jiang, C., Wang, J. & Liu, S. Analyzing temperature and precipitation extremes in China using multiple gridded datasets: A comparative evaluation. *Weather Clim. Extremes* **42**, 100614. https://doi.org/10.1016/j.wace.2023.100614 (2023).
5. Yin, H., Donat, M. G., Alexander, L. V. & Sun, Y. Multi-dataset comparison of gridded observed temperature and precipitation extremes over China. *Int. J. Climatol.* **35**(10), 2809–2827. https://doi.org/10.1002/joc.4174 (2015).
6. Hofstra, N., New, M. & McSweeney, C. The influence of interpolation and station network density on the distributions and trends of climate variables in gridded daily data. *Clim. Dyn.* **35**(5), 841–858. https://doi.org/10.1007/s00382-009-0698-1 (2010).
7. Gross, M. H., Donat, M. G., Alexander, L. V. & Sisson, S. A. The sensitivity of daily temperature variability and extremes to dataset choice. *J. Clim.* **31**(4), 1337–1359. https://doi.org/10.1175/JCLI-D-17-0243.1 (2018).
8. Pitman, A. J. & Perkins, S. E. Global and regional comparison of daily 2-m and 1000-hPa maximum and minimum temperatures in three global reanalyses. *J. Clim.* **22**(17), 4667–4681. https://doi.org/10.1175/2009JCLI2799.1 (2009).
9. Thorne, P. W. & Vose, R. S. Reanalyses suitable for characterizing Long-Term trends. *Bull. Am. Meteorol. Soc.* **91**(3), 353–362. https://doi.org/10.1175/2009BAMS2858.1 (2010).
10. de Coughlan, E., Arrighi, J. & Marunye, J. Challenging the universality of heatwave definitions: Gridded temperature discrepancies across climate regions. *Clim. Change* **176**(12), 167. https://doi.org/10.1007/s10584-023-03641-x (2023).
11. Lader, R., Bhatt, U. S., Walsh, J. E., Rupp, T. S. & Bieniek, P. A. Two-Meter temperature and precipitation from atmospheric reanalysis evaluated for Alaska. *J. Appl. Meteorol. Climatology* **55**(4), 901–922. https://doi.org/10.1175/JAMC-D-15-0162.1 (2016).
12. Dunn, R. J. H. et al. Development of an updated global land in Situ-Based data set of temperature and precipitation extremes: HadEX3. *J. Geophys. Research: Atmos.* **125**(16). https://doi.org/10.1029/2019JD032263 (2020). e2019JD032263.
13. Hersbach, H. et al. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **146**(730), 1999–2049. https://doi.org/10.1002/qj.3803 (2020).
14. Muñoz-Sabater, J. et al. ERA5-Land: Astate-of-the-art global reanalysis dataset for land applications. *Earth Syst. Sci. Data.* **13**(9), 4349–4383. https://doi.org/10.5194/essd-13-4349-2021 (2021).
15. Daly, C. et al. Physiographically sensitive mapping of Climatological temperature and precipitation across the conterminous united States. *Int. J. Climatol.* **28**(15), 2031–2064. https://doi.org/10.1002/joc.1688 (2008).
16. Gelaro, R. et al. The Modern-Era retrospective analysis for research and applications, version 2 (MERRA-2). *J. Clim.* **30**(14), 5419–5454. https://doi.org/10.1175/JCLI-D-16-0758.1 (2017).
17. Smith, E., Rose, S., Chiang, F., Diaz, D. & Fischer, L. *Developing Local Climate Change Information: Steps and Illustrative Analysis* (EPRI, 2023).
18. Peng, H., Wang, S. & Wang, X. Consistency and asymptotic distribution of the Theil–Sen estimator. *J. Stat. Plann. Inference* **138**(6), 1836–1850 (2008).
19. Sheridan, S. C., Lee, C. C. & Smith, E. T. A comparison between station observations and reanalysis data in the identification of extreme temperature events. *Geophys. Res. Lett.* **47**(15), e2020GL088120. https://doi.org/10.1029/2020GL088120 (2020).
20. You, Q. et al. Can temperature extremes in China be calculated from reanalysis? *Glob. Planet Change.* **111**, 268–279. https://doi.org/10.1016/j.gloplacha.2013.10.003 (2013).
21. Rasmussen, R. M. et al. CONUS404: the NCAR–USGS 4-km long-term regional hydroclimate reanalysis over the CONUS. *Bull. Am. Meteorol. Soc.* **104**(8), E1382–E1408 (2023).

Acknowledgements

Author contributions

E.S. was the principal investigator, designed the analysis, prepared the figures, and contributed heavily to the writing of the manuscript. N.G. helped lead the writing of the manuscript and contributed to the analysis. X.L. led the analysis for the manuscript.D.D. provided technical support for the analysis and review of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-97570-7.

Correspondence and requests for materials should be addressed to E.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.