A comparison between station observations and reanalysis data in the identification of extreme temperature events

Scott C. Sheridan, Cameron C. Lee, Erik T. Smith

Department of Geography, Kent State University, Kent, OH 44242 USA

Abstract

While many studies comparing atmospheric reanalysis and surface observations have focused on the similarity of mean fields, trends, or frequencies of extreme events, very few have assessed how similar surface observations and reanalysis data sets are in terms of their specific identification of extreme temperature event days. Here, we assess the similarity between surface observations and three reanalysis products: ERA5, ERA5-LAND, and NARR, in terms of the days on which they identify extreme temperature events. We assess similarity from 1979-2016 for 231 locations in the United States and Canada, assessing Extreme Heat and Cold Event days, as well as their counterpart events that are relative for the time of year. Cold Events have a greater match than Heat Events. ERA5 has the greatest match percentage with station data across the study region. Match percentage is greatest in mid-latitude, continental locations, with poorer performance in coastal areas, and the Arctic.

Plain Language Summary

Atmospheric reanalysis products are simulations of the atmosphere that create gridded historical data sets. While many studies have looked at how well these products match surface observations overall, very few have examined how similar they are in identifying extreme weather event days. Here, we explore this similarity using three reanalysis products: ERA5, ERA5-LAND, and NARR, in terms of the days on which they identify extreme temperature events. We do this for 231 locations in the United States and Canada from 1979-2016, for both Extreme Heat and Cold. Cold Events have a greater match than Heat Events. ERA5 has the greatest match percentage with station data across the study region. Match percentage is greatest in mid-latitude, continental locations, with poorer performance in coastal areas, and the Arctic.

Keywords: extreme temperature events, heat waves, cold waves, reanalysis, North America

1 Introduction

Over the past 25 years, as the availability of atmospheric reanalysis data sets has proliferated, their prominence in climate research has grown considerably. These reanalysis products, in which observed data from a number of platforms are assimilated and then gridded through short-term simulations within a modern forecast system, have much to offer climate science – the data sets are complete, available for many atmospheric variables, produced on a regular grid at many vertical levels, can have extremely fine-scale spatial resolution, and comprise multiple decades (Dee et al., 2014). All of these facets greatly standardize and enhance spatio-temporal analyses, and thus it is unsurprising that these data sets are frequently used in climate change and variability research. Indeed, for many climate change and variability applications, such as the global energy balance (Trenberth et al., 2011), global reanalysis products are a necessity.

Nevertheless, there are marked differences between reanalysis and observation data sets. Some of this is by design; for instance, the NCEP-NCAR reanalysis (NNR; Kalnay et al. 1996) intentionally does not incorporate surface temperature observations, as a means of potentially identifying the influence of urbanization or other land use changes have on the climate system (Cornes and Jones, 2013). Alternatively, the JRA-55 reanalysis eschews satellite data to provide an historical reconstruction more consistent in terms of input variables (Kobayashi et al., 2015). There is a general understanding that some of the variables must be used with greater caution, particularly those related to the hydrologic cycle and in areas with sparse surface data, nevertheless these reanalysis products are still valuable (Essou et al., 2016).

There is often an assumption made that surface observation data and reanalysis data are one and the same, or at least similar enough for most comparative or integrative purposes, although this is not without contention (Parker, 2016). A number of studies as a result have focused on confirming that reanalysis and observation data sets have similar climatologies. (e.g. Schoof et al., 2017; Kishore et al., 2016; Behrangi et al., 2016), although nearly all studies identify regional biases or distinctions when multiple reanalysis data sets are compared.

Extreme events are the most critical aspect of many applied climatological studies as they can be harmful to human and natural systems. In particular, heat and cold events have been linked to anomalous human mortality (e.g., Sheridan and Allen, 2018; Smith and Sheridan, 2019), agricultural losses (e.g., Teixera et al., 2013), infrastructure damage (e.g. Xia et al., 2013) and other detrimental effects. In turn, studies comparing reanalysis and observational data sets have explored differences in the climatologies of such events, some downscaling to the individual station level. Broadly, there is a strong correlation between temperatures derived from reanalysis and their station counterparts, yet there can be substantial bias that impacts the climatology of extreme temperature days (Lader et al., 2016). Some studies have shown a general alignment in trends between data sets, such as for the US in Schoof et al. (2017). However, in other studies, trends in the most extreme hot days have the poorest alignment among the reanalysis data sets themselves (Pitman and Perkins, 2009), as well as with observations (Europe; Cornes and Jones, 2013; Africa; Ceccherini et al., 2015). Over China, You et al. (2013) showed that patterns of cold-related variables were reproduced most poorly. Within all temperature reanalysis products, complex geography can make it difficult to resolve precise temperature values for locations due to the scale of reanalysis being inconsistent with many physical climate processes (Holden et al., 2016).

Nevertheless, for many applications, such as heat-related mortality, the identification of specific events is critical – namely, do reanalysis and observation data sets identify the same heat events as occurring? This is indispensable for research, but not very well understood. We only know of one previous work that examined alignment between observations and reanalysis of specific extreme temperature event identification (Ceccherini et al., 2015), and another that assessed the differences in the weather-human mortality relationship between station data and gridded interpolations (Spangler et al., 2019). Thus, in this paper, we directly address this question by analyzing the similarity between extreme temperature event identification within four data sets: surface synoptic weather observations (SYNOP) in the United States and Canada, the newly released ECMWF-based reanalysis (ERA5, Hersbach et al., 2019) – which does assimilate surface temperature observations and has been shown to reduce temperature bias from the previous generation ERA-Interim (Betts, Chan, & Desjardins, 2019; Johannsen et al., 2019), the ERA5-LAND reanalysis, which offers enhanced resolution compared to ERA5 itself but does not directly assimilate surface observations, and the North American Regional Reanalysis (NARR; Mesinger et al., 2006), which also does not use surface observations in its assimilation scheme. The comparisons are made across 230 locations in the United States and Canada for the period 1979-2016. We address similarities in the climatology, as well as the identification of specific extreme temperature events (ETE) among the data sets.

2 Data and methods

2.1 Meteorological Data

Station-observation data for the 231 stations in North America were obtained from the National Center of Environmental Information (NCEI) and Environment Canada for all sites in Table S1. The threshold for inclusion of this study was at least 97.5% completeness in the station observation record.

In order to calculate daily-scale apparent temperatures, 2-m temperature, 2-m dew points, and 10-m zonal and meridional components of the wind were obtained from the National Centers for Environmental Prediction (NCEP) North American Regional Reanalysis (NARR) data set and the ERA5 and ERA5-LAND reanalysis data sets from the European Centre for Medium-Range Weather Forecast for 1979-2016 (1981-2016 for ERA5-LAND). These data were acquired for the nearest land-based grid point in each reanalysis domain to each of the 231 surface weather stations. The spatial resolution of the NARR (~32km) and ERA5 (~30km) are very similar, thus complex geography, such as mountains or water bodies, are likely to have a consistent impact on both data sets and were not therefore considered when finding the nearest grid-point. The native resolution of ERA5-LAND is ~9km.

The study period of 1979-2016 is entirely within the satellite era to remove potential biases (Jones et al., 2012). We also acknowledge that other studies have used gridded surface data (Schoof et al., 2017) such as PRISM (Daly et al., 2008), that attempt to reconcile station inhomogeneities that are present in many data sets (e.g., Brown and DeGaetano, 2013); however, for many applied climate studies, particularly point-based ones, station data are still used, and hence we incorporate the historical station record itself.

Due to the variable amounts of atmospheric moisture associated with heat events in North America, most assessments of heat events typically use an apparent temperature index accordingly (McGregor and Vanos, 2018), and apparent temperature metrics are used in official threshold delineation in the US (Weinberger et al., 2018) and Canada (Benmarhnia et al., 2016). Thus, for each of the three data sets, 2-m temperature, 2-m dew point, and 10-m wind speed were obtained/calculated for the eight 3-hourly observations per day, a temporal time frame with improved results over 12-hourly values (Cornes and Jones, 2013). For each of these three-hourly observations, an apparent temperature was calculated based on the Steadman (1984) formula for outdoor shade conditions:

$$AT = -2.7 + 1.04T + 2.0P - 0.65u;$$

in which *T* and *AT* are temperature and apparent temperature in °C, *P* is vapor pressure in kPa (calculated from dew points in °C), and *u* is wind speed in m/s. A daily mean apparent temperature (AT) is then calculated from the 8 observations per day. To align the definition of a 'local day' across the continent, centered on midnight to midnight, for each day the observations of 0900, 1200, 1500, 1800, 2100, 0000 (+1-day), 0300 (+1-day), and 0600 (+1-day) GMT were used. These times equate to 0100 to 2200 Pacific Standard Time, and 0400 to 0100 Eastern Standard Time. This daily mean AT is the basis for all subsequent calculations of ETEs.

2.2 Identification of Extreme Heat and Cold Events

As many studies of ETE are based upon their human impact, in this paper Extreme Heat Events (EHE), Extreme Cold Events (ECE), Relative Extreme Heat Events (REHE), and Relative Extreme Cold Events (RECE) are based upon the initial work by Nairn and Fawcett (2014) defining EHE, and later adaptations by the authors of this article (Sheridan and Lee, 2019; Sheridan et al., 2019), in which a connection between this definition of ETE and mortality has been shown.

The EHE is initially based on the Excess Heat Factor (EHF), defined as the product of the magnitude of a heat event and an acclimatization term. The magnitude of the heat event (excess heat, EH) is calculated as:

$$EH = \max\left(0, \left(\textstyle\sum_{i=-2}^{0} AT_i\right)/3 - AT_{95}\right), \quad (1)$$

where AT_i is the apparent temperature on day *i*, averaged over a three-day period, and AT_{95} is the overall 95th percentile of daily mean apparent temperature for a location (based on the 1981-2010 normal period). It should be noted that this percentile is calculated separately for each of the three data sets to reduce systematic bias in the variables.

The acclimatization term is:

$$EH_{accl} = \left(\textstyle\sum_{i=-2}^{0} AT_i\right)/3 - \left(\textstyle\sum_{i=-32}^{-3} AT_i\right)/30, \quad (2)$$

representative of the difference between the three-day mean apparent temperature and the mean of the 30 days prior. This is critical to the Nairn and Fawcett (2014) methodology as it addresses the increased vulnerability to heat when there has been a lack of short-term acclimatization, something identified in literature (e.g., Lee et al., 2014).

EHF then is the product of these two terms:

$$EHF = \max(0, EH) \times \max(1, EH_{accl}), \quad (3)$$

in units of K^2. To define an extreme heat event (EHE), we use the Nairn and Fawcett (2014) definition, whereby the *EHF* must exceed the 85th percentile of all positive *EHF* values for a location over the climatological period.

The concept of Extreme Cold Event (ECE) identification is similar, except with the 5th percentile of apparent temperature (AT_5) as the basis for excess cold (EC) being identified, and the 15th percentile threshold of ECF used to identify ECE days, with excess cold factor defined as:

$$ECF = -1 \times \min(0, EC) \times \min(-1, EC_{accl}). \quad (4)$$

In contrast to absolute heat events and cold events, events that are extreme relative to the time of year are also of interest (Sheridan and Lee, 2019). We thus also assess relative EHF (REHF) and ECF (RECF). The definitions are similar to the EHF and ECF above, except that these two variables use a percentile threshold that varies seasonally, calculated as the 92.5th (REHF)/7.5th (RECF) percentile over the climatological period for the 15 days centered on the day being evaluated. Relative Extreme Heat Events (REHE) are identified as all days above the 85th percentile distribution of REHF, and relative extreme cold events (RECE) as days below the 15th percentile distribution of RECF.

2.3 Match of days identified as EHE and ECE

In addition to the overall similarity of trends, a comparison is made between the similarity of the exact days that are identified as EHE, ECE, REHE, and RECE within the 4 data sets. Of principal interest is the match between each of the three reanalysis data sets and the station data, and so our primary calculation is the match percentage defined as the percent of ETE days that are identified in the station data that are also identified in each reanalysis data set. To assess the identification more broadly, we also calculate the match percentage when there is any extreme temperature factor, as defined above.

3 Results

Table 1 shows the overall sample size of ETE. Each data set identifies between 2.2-2.8 ETE events across the study region, with 14.4-18.7 days/year identified as having some extreme temperature factor. It should be noted that, while each data set’s thresholds are separately identified by using the same percentage thresholds, values will not be identical across data sets due to the multiple-day component of ETE definition. Thus, there are a greater number of days identified in areas that ETEs persist longer.

Further, the definitions are defined using the 1981-2010 normal period, and with REHE in particular, the substantive increase in the 2010s increases the overall sample size.

The spatial pattern of ETE is shown in Figure 1. ECE tend to be most frequent in the midwestern US and Rocky Mountains, and least frequent in the southern US and other coastal regions. EHE are most frequent across the southern tier of the US, and least frequent farther north. The frequency of RECE is roughly equal everywhere except eastern Canada and the northeastern US, where it is less frequent. REHE also roughly equally common everywhere, except in some areas of the High Plains. The spatial patterns of ETE as defined by the three reanalysis products are all similar to those of the station data, although there is an overestimation of most ETE across the midwestern US, particularly with NARR.

Of the event types, ECE have the greatest correlations between station and reanalysis data sets (Table 1; Figures 2-4), with the overall match ranging from 72% with NARR to 74% with ERA5-LAND and 81% with ERA5. Days with any ECF identified range in match percentage from 81% to 89%. Spatially, there is considerable variability – station-defined ECEs are best matched across the more continental locations with little topography, peaking with a 98.6% match percentage between ERA5 and station ECE days at Montgomery, Alabama. The similarity between ETEs extends all the way to the Gulf and Atlantic coast, where ECEs generally arrive from the north and thus coastal interactions would be minimized. There is a notable drop in match percentage at several stations downwind of the Great Lakes, such as Buffalo, New York, and Erie, Pennsylvania, suggesting that lake-induced air mass-modification may not be well captured in ECEs.

A much greater variability in match percentage occurs across the topographically varied terrain of the western half of Canada and the US. At some sites where extreme cold air masses have a very specific trajectory, such as the immediate coastal cities of Prince Rupert, British Columbia and Quillayute, Washington, there is strong agreement, while at others where the coastal plain is extremely narrow (e.g. Arcata, California), match percentage is much lower. The weakest agreement is at the northernmost stations, generally inland or Arctic-facing locations north of 60°N, where all reanalysis products struggle to simulate the coldest days. Match values are below 20% at several stations; the NARR data set identifies only 25% of the excess cold factor days that are identified by the station data set.

The ERA5 reanalysis performs better for ECE than NARR by 9 percentage points overall (Figure S1). Most stations are better simulated by ERA5, with the greatest improvement seen at some western stations as well as some, but not all, locations around 60°N. The ERA5-LAND match percentage is generally in between that of the other two datasets, generally mimicking ERA5 itself spatially but with slightly worse match throughout the study region. However, in some individual cases, the match is considerably different; for instance, ERA5-LAND is by far the worst in terms of station-data match at Traverse City, Michigan, but consistently best at all Alaskan stations on the Bering Sea coast.

EHEs, in contrast, do not show as large of an association between the reanalysis and the station data sets, in particular for NARR, for which there is only a 57% match overall. The relationship is quite variable across space, with once again the peak similarity observed in the most continental locations, albeit shifted rather north from the ECE peak, with the absolute highest match (91.3%) in Kaspuskaing, Ontario, and match percentages above 80% north through the Arctic. The matches are considerably weaker in the southern and eastern US, where the contribution of humidity to the overall apparent temperature would be greatest, and thus any discrepancies in resolution of high dew points may be magnified. The ability to simulate EHEs is especially poor in the extreme southern regions of the study where summertime thermal variability is low, with only a 40-50% match at Miami, Florida, and a 21-31% match at Key West, Florida.

Across the western half of the study region, there is once again considerable spatial variability with EHE; many of the locations in the Great Basin and intermountain west are very well simulated, although there are a number of outliers. Across the Pacific coast, match percentages tend to drop precipitously, with the lowest values at stations along the Pacific Coast, reaching as low as a 10% match between stations-data and NARR EHE days in Arcata, California.

The difference between the ERA5 and the NARR is greater with EHE than it was with ECE, with a 15-percentage-point difference overall; ERA5-LAND again is roughly halfway in between. These differences are greatest across several different regions of the east central US, but are most pronounced across the rapidly urbanizing southwestern US cities, likely a result of the different data assimilation schemes. For example, in Phoenix, Arizona and Las Vegas, Nevada, with their substantive and complex heat islands (Wang et al., 2018), ERA5 has 76% and 86% match with station data, respectively, compared to only 41% and 56% with the NARR data set, and 65% and 69% with ERA5-LAND.

For the relative events, RECE and REHE, the patterns are broadly similar to their absolute counterparts, ECE and EHE. Interestingly, there is an overall modest improvement in REHE match compared to EHE, whereas RECE show a slightly lower correspondence among the data sets. In comparing the NARR to ERA5, the NARR REHE match drops substantively in the Canadian Rockies compared to the EHE, whereas with the ERA5 the match percentages for EHE and REHE here are relatively similar. Conversely, across the midwestern US, NARR improves considerably in match with REHE identification, to the point where it is similar in skill to ERA5.

Across all data sets, there is no statistically significant improvement in terms of match percentage over time.

4 Discussion and conclusions

In this research, we have shown that, while all reanalysis data sets tested broadly replicate the spatial pattern of extreme temperature events, there is a clear discrepancy of which days are actually identified. Coastal locations show the greatest discrepancy with station observations among reanalysis data sets, though we noticed no clear difference in match based on level of urbanization, aside from the very rapidly urbanizing areas of the desert southwestern US. These results are similar to those of Ceccherini et al. (2015), who examined the match across Africa, although their work only evaluated one reanalysis (ERA-INTERIM). The ERA5, which is the only one of the three reanalysis data sets that directly integrates surface observations, was the best performing reanalysis data set. The indirect data assimilation of surface observations by the ERA5-LAND and lack of atmosphere and ocean coupling (Yang and Sabater, 2020) may explain why the ERA5-LAND has generally lower match percentages with station observations than the ERA5. While the ERA5-LAND has a higher spatial resolution than the NARR, it also benefits from newer bias correction and parameterization schemes, thus it is difficult to determine whether the improvement of the ERA5-LAND over the NARR is more attributed to model physics or spatial resolution. However, the difference between the ERA5-LAND and the NARR is largest across geographically diverse regions such as western North America, with several ERA5-LAND locations along the coast of Alaska having higher skill than the ERA5. This suggests that not only are improvements in data assimilation important, but higher spatial resolution is critical to accurately reproducing extreme events observed from surface weather stations.

We acknowledge that, of course, the surface-observation data set is not without bias itself; there are a number of potential discontinuities due to equipment changes over time (Guttman and Baker, 1996), and there are trends due to urbanization (which may or may not be desired, based on application; e.g. Jin et al., 2018). However, data observed at airports represent the most complete set of historical meteorological observations and are often used as the 'reference' data set in climatological research, as they are herein. While some new research has shown that the weather-human health relationship can be successfully simulated using gridded interpolations (Spangler et al., 2019) from reconstructed data sets, nevertheless, it is quite likely that the use of a single observation site to represent a broad area will still be used moving forward. For some application studies, e.g. human health, there has been a greater emphasis found on finding the 'best' exposure metric (Anderson et al., 2013) than studies evaluating the appropriateness of the site or data source, such as evaluating the selection of which station to use (de'Donato et al., 2018) or the relationship between indoor and outdoor conditions (e.g, Nguyen and Dockery, 2016). Given many applied studies require event identification as a the fundamental starting point, while there have been many studies evaluating how well data sets align in terms of climatology or trends (e.g., Schoof et al., 2017; Cornes et al., 2013), there needs to be considerably more research on how well the identified events themselves match across data sets.

Data Availability

Binary data sets of heat event identification will be uploaded to Mendeley and are attached as supported information for the review process.

References

Anderson, G.B., Bell, M.L. and Peng, R.D., 2013. Methods to calculate the heat index as an exposure metric in environmental health research. *Environmental health perspectives*, 121(10), 1111-1119.

Behrangi, A., Christensen, M., Richardson, M., Lebsock, M., Stephens, G., Huffman, G.J., Bolvin, D., Adler, R.F., Gardner, A., Lambrigtsen, B. and Fetzer, E., 2016. Status of high-latitude precipitation estimates from observations and reanalyses. Journal of Geophysical Research: *Atmospheres*, 121(9), 4468-4486.

Benmarhnia, T., Bailey, Z., Kaiser, D., Auger, N., King, N. and Kaufman, J.S., 2016. A difference-in-differences approach to assess the effect of a heat action plan on heat-related mortality, and differences in effectiveness according to sex, age, and socioeconomic status (Montreal, Quebec). *Environmental health perspectives,* 124(11), 1694-1699.

Betts, A. K., Chan, D. Z., and Desjardins, R. L., 2019. Near-surface biases in ERA5 over the Canadian Prairies. *Frontiers in Environmental Science*, 7, 129.

Brown, P.J. and DeGaetano, A.T., 2013. Trends in U.S. Surface Humidity, 1930–2010. *Journal of Applied Meteorology and Climatology*, 52, 147–163.

Ceccherini, G., Russo, S., Ameztoy, I., Marchese, A.F. and Carmona-Moreno, C., 2017. Heat waves in Africa 1981-2015, observations and reanalysis. *Natural Hazards & Earth System Sciences*, 17(1).

Cornes, R. C., and Jones, P. D., 2013. How well does the ERA-Interim reanalysis replicate trends in extremes of surface temperature across Europe? *Journal of Geophysical Research Atmospheres*, 118, 10,262–10,276, doi:10.1002/jgrd.50799.

Daly, C., M. Halbleib, J. I. Smith, W. P. Gibson, M. K. Doggett, G. H. Taylor, J. Curtis, and P. P. Pasteris, 2008. Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. Int. J. Climatol., 28, 2031–2064, doi:10.1002/joc.1688.

De'Donato, F.K., Stafoggia, M., Rognoni, M., Poncino, S., Caranci, N., Bisanti, L., Demaria, M., Forastiere, F., Michelozzi, P., Pelosini, R. and Perucci, C.A., 2008. Airport and city-centre temperatures in the evaluation of the association between heat and mortality. *International Journal of Biometeorology*, 52(4), 301-310.

Dee, D.P., Balmaseda, M., Balsamo, G., Engelen, R., Simmons, A.J. and Thépaut, J.N., 2014. Toward a consistent reanalysis of the climate system. *Bulletin of the American Meteorological Society*, 95(8), 1235-1248.

Essou, G.R., Sabarly, F., Lucas-Picher, P., Brissette, F. and Poulin, A., 2016. Can precipitation and temperature from meteorological reanalyses be used for hydrological modeling?. *Journal of Hydrometeorology*, 17(7), 1929-1950.

Guttman, N.B. and Baker, C.B., 1996. Exploratory analysis of the difference between temperature observations recorded by ASOS and conventional methods. *Bulletin of the American Meteorological Society*, 77(12), 2865-2874.

Hersbach, H., Bell, B., Berrisford, P., Horányi, A., Sabater, J.M., Nicolas, J., Radu, R., Schepers, D., Simmons, A., Soci, C. and Dee, D. 2019. Global reanalysis: goodbye ERA-Interim, hello ERA5. *ECMWF Newsletter*, 159, 17-24.

Holden, Z.A., Swanson, A., Klene, A.E., Abatzoglou, J.T., Dobrowski, S.Z., Cushman, S.A., Squires, J., Moisen, G.G. and Oyler, J.W., 2016. Development of high-resolution (250 m) historical daily gridded air temperature data using reanalysis and distributed sensor networks for the US Northern Rocky Mountains. *International Journal of Climatology*, 36(10), pp.3620-3632.

Jin, K., Wang, F., Yu, Q., Gou, J. and Liu, H., 2018. Varied degrees of urbanization effects on observed surface air temperature trends in China. *Climate Research*, 76(2), 131-143.

Johannsen, F., Ermida, S., Martins, J., Trigo, I. F., Nogueira, M., & Dutra, E., 2019. Cold Bias of ERA5 Summertime Daily Maximum Land Surface Temperature over Iberian Peninsula. *Remote Sensing*, 11(21), 2570.

Jones, P. D., Lister, D. H., Osborn, T. J., Harpham, C., Salmon, M., and Morice, C. P., 2012. Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010. *Journal of Geophysical Research*, 117, D05127, doi:10.1029/2011JD017139.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J. and Zhu, Y., 1996. The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, 77(3), 437-472.

Kishore, P., Jyothi, S., Basha, G., Rao, S.V.B., Rajeevan, M., Velicogna, I. and Sutterley, T.C., 2016. Precipitation climatology over India: validation with observations and reanalysis datasets and spatial trends. *Climate dynamics*, 46(1-2), 541-556.

Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., Onogi, K., Kamahori, H., Kobayashi, C., Endo, H. and Miyaoka, K., 2015. The JRA-55 reanalysis: General specifications and basic characteristics. *Journal of the Meteorological Society of Japan,* Ser. II, 93(1), 5-48.

Lader, R., Bhatt, U.S., Walsh, J.E., Rupp, T.S. and Bieniek, P.A., 2016. Two-meter temperature and precipitation from atmospheric reanalysis evaluated for Alaska. *Journal of Applied Meteorology and Climatology*, 55(4), 901-922.

Lee, M., Nordio, F., Zanobetti, A., Kinney, P., Vautard, R. and Schwartz, J., 2014. Acclimatization across space and time in the effects of temperature on mortality: a time-series analysis. *Environmental Health*, 13(1), 89.

McGregor, G.R. and Vanos, J.K., 2018. Heat: a primer for public health researchers. *Public health*, 161, 138-146.

Mesinger, F., DiMego, G., Kalnay, E., Mitchell, K., Shafran, P.C., Ebisuzaki, W., Jović, D., Woollen, J., Rogers, E., Berbery, E.H. and Ek, M.B., 2006. North American regional reanalysis. *Bulletin of the American Meteorological Society*, 87(3), 343-360.

Nairn, J. R. and Fawcett, R. J., 2014. The excess heat factor: a metric for heatwave intensity and its use in classifying heatwave severity. *International Journal of Environmental Research and Public Health*, 12(1), 227-253.

Nguyen, J.L. and Dockery, D.W., 2016. Daily indoor-to-outdoor temperature and humidity relationships: a sample across seasons and diverse climatic regions. *International journal of biometeorology*, 60(2), pp.221-229.

Parker, W.S., 2016. Reanalyses and observations: What's the difference?. *Bulletin of the American Meteorological* Society, 97(9), 1565-1572.

Perkins, S.E. and Alexander, L.V., 2013. On the measurement of heat waves. *Journal of Climate*, 26(13), 4500-4517.

Pitman, A.J. and S.E. Perkins, 2009. Global and Regional Comparison of Daily 2-m and 1000-hPa Maximum and Minimum Temperatures in Three Global Reanalyses. *J. Climate*, 22, 4667–4681.

Schoof, J.T., Ford, T.W. and Pryor, S.C., 2017. Recent changes in US regional heat wave characteristics in observations and reanalyses. *Journal of Applied Meteorology and Climatology*, 56(9), 2621-2636.

Sheridan, S.C., and Allen, M.J., 2018. Temporal trends in human vulnerability to excessive heat. *Environmental Research Letters*, 13, 043001, 12 pp.

Sheridan, S.C., Lee, C.C., and Allen, M.J., 2019. The mortality response to absolute and relative temperature extremes. *International Journal of Environmental Research and Public Health*, 16, 1493.

Sheridan, S.C., and Lee, C.C., 2019. Temporal trends in absolute and relative extreme temperature events across North America. *Journal of Geophysical Research – Atmospheres*, 123, 11889-11898.

Smith, E.T., and Sheridan, S.C., 2019. The Influence of Extreme Cold Events on Mortality in the United States. *Science of the Total Environment*, 67, 342-251.

Spangler K.R., Weinberger K.R., and Wellenius, G.A., 2019. Suitability of gridded data sets for use in environmental epidemiology. *Journal of Exposure Science and Environmental Epidemiology*, 29, 777-789.

Steadman, R. G. (1984). A universal scale of apparent temperature. *Journal of Climate and Applied Meteorology,* 23(12), 1674-1687.

Teixeira, E.I., Fischer, G., Van Velthuizen, H., Walter, C. and Ewert, F., 2013. Global hot-spots of heat stress on agricultural crops due to climate change. *Agricultural and Forest Meteorology*, 170, 206-215.

Trenberth, K.E., Fasullo, J.T. and Mackaro, J., 2011. Atmospheric moisture transports from ocean to land and global energy flows in reanalyses. *Journal of climate*, 24(18), 4907-4924.

Wang, C., Middel, A., Myint, S.W., Kaplan, S., Brazel, A.J. and Lukasczyk, J., 2018. Assessing local climate zones in arid cities: The case of Phoenix, Arizona and Las Vegas, Nevada. *ISPRS Journal of Photogrammetry and Remote Sensing*, 141, 59-71.

Weinberger, K.R., Zanobetti, A., Schwartz, J. and Wellenius, G.A., 2018. Effectiveness of National Weather Service heat alerts in preventing mortality in 20 US cities. *Environment international*, 116, 30-38.

Xia, Y., Van Ommeren, J.N., Rietveld, P. and Verhagen, W., 2013. Railway infrastructure disturbances and train operator performance: The role of weather. *Transportation research part D: transport and environment*, 18, 97-102.

Yang, X., and J., M., Sabater. (2020, March 13). “ERA5-Land: Data Documentation - Copernicus Knowledge Base.” Retrieved from: https://confluence.ecmwf.int/display/CKB/ERA5-Land%3A+data+documentation#ERA5Land:datadocumentation-References.

You Q., Fraedrich, K., Min, J., Kang, S., Zhu X., Ren, G., Meng, X., 2013. Can temperature extremes in China be calculated from reanalysis? *Global and Planetary Change*, 111, 268-279.

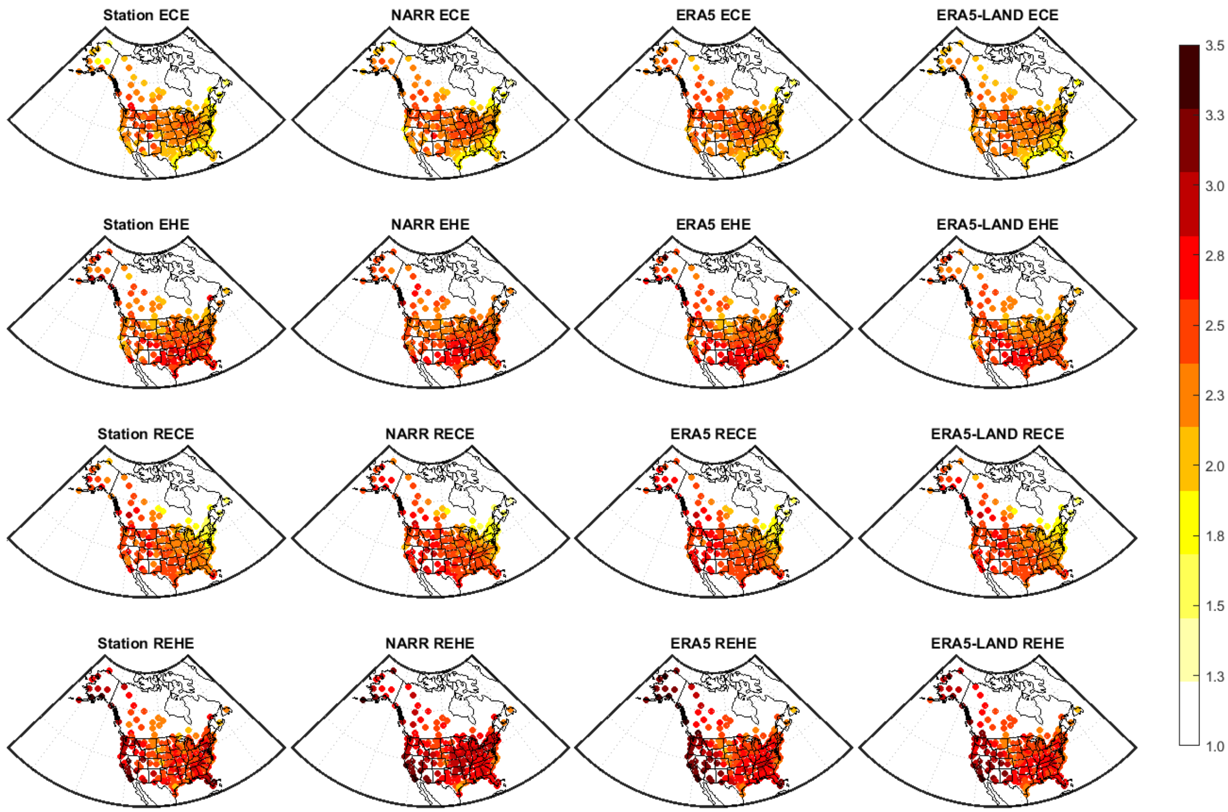

Figure 1. Mean annual number of days classified as ECE (extreme cold event), EHE (extreme heat event), RECE (relative extreme cold event), and REHE (relative extreme heat event) for each of the four data sets of the study.

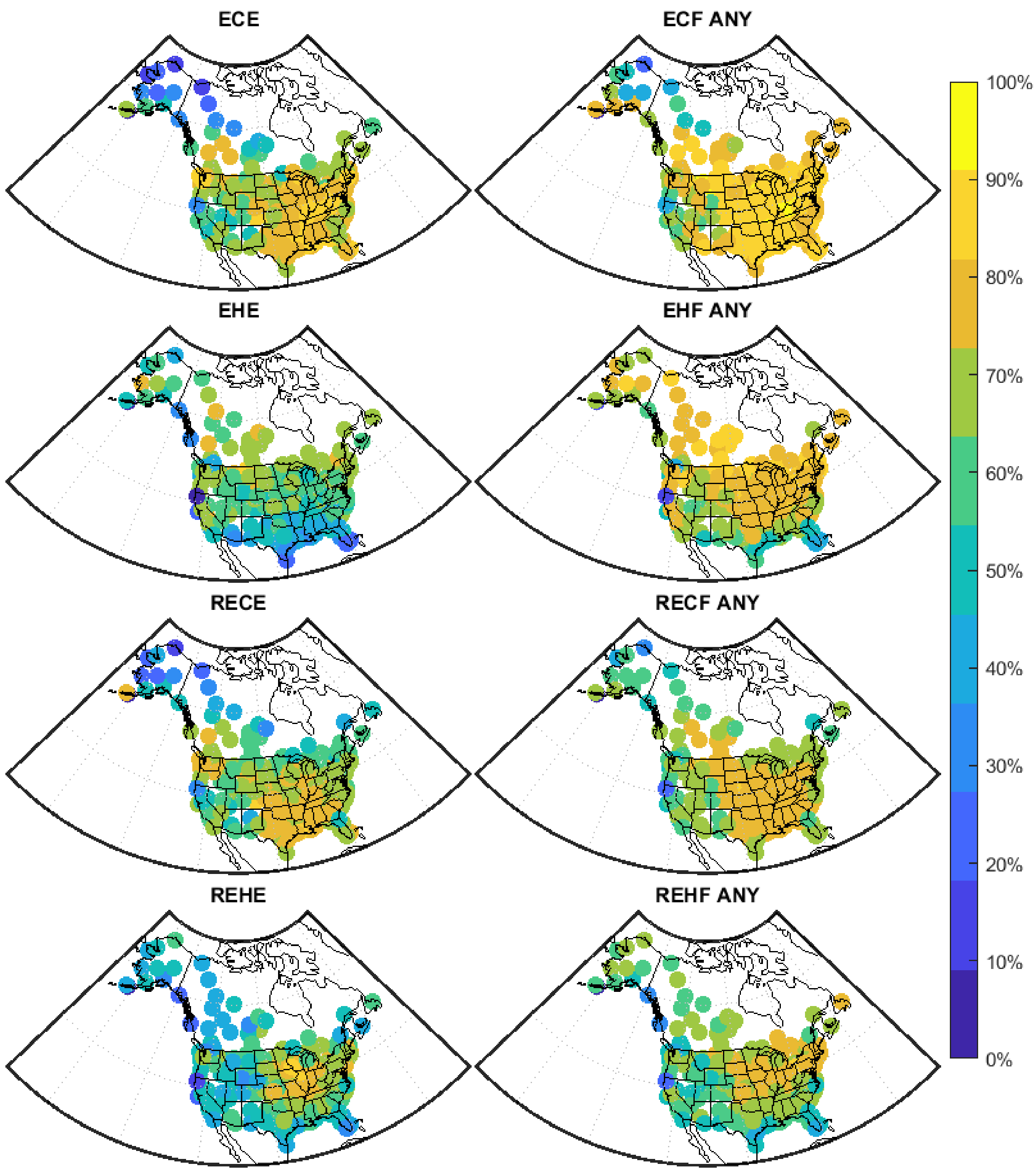

Figure 2. ETE match percentage between station-defined events and those defined using NARR. The left column compares days that are identified as being events (ECE, EHE, RECE, REHE) while the right column compares days in which each excess factor is non-zero (ECF, excess cold factor; EHF, excess heat factor; RECF, relative excess cold factor; REHF, relative excess heat factor).

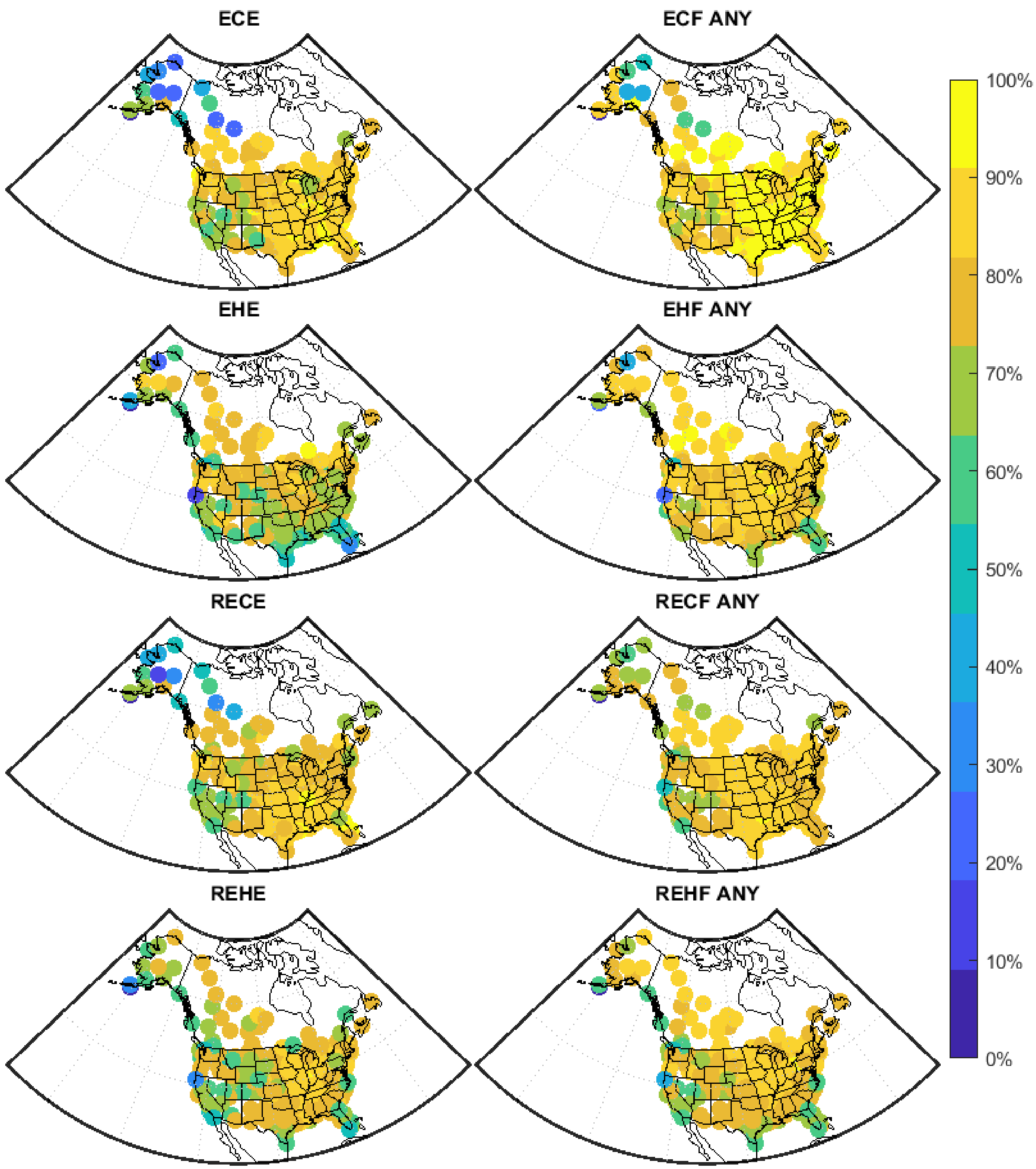

Figure 3. Same as Figure 2 except for comparison between station-defined events and ERA5.

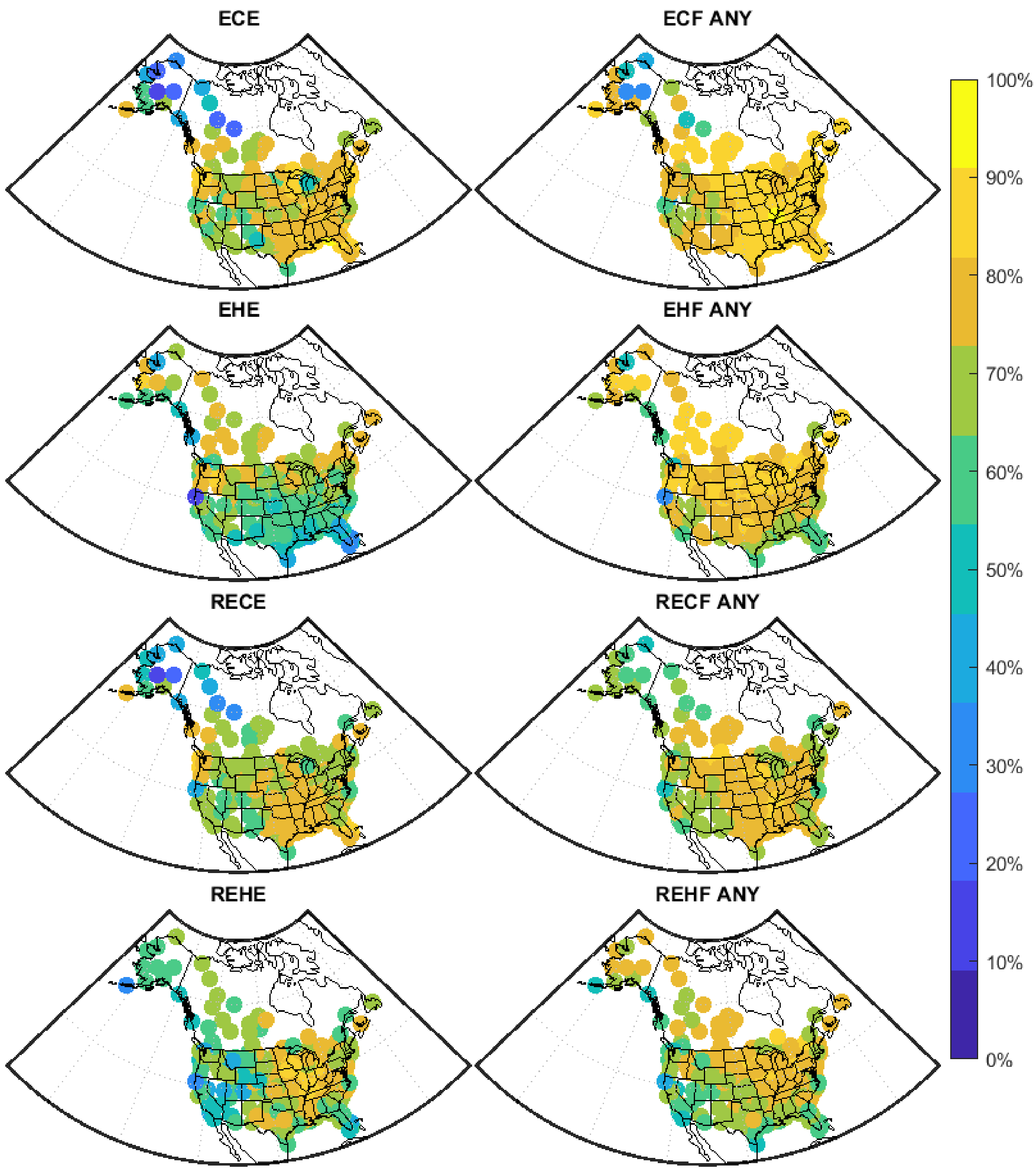

Figure 4. Same as Figure 2 except for comparison between station-defined events and ERA5-LAND.

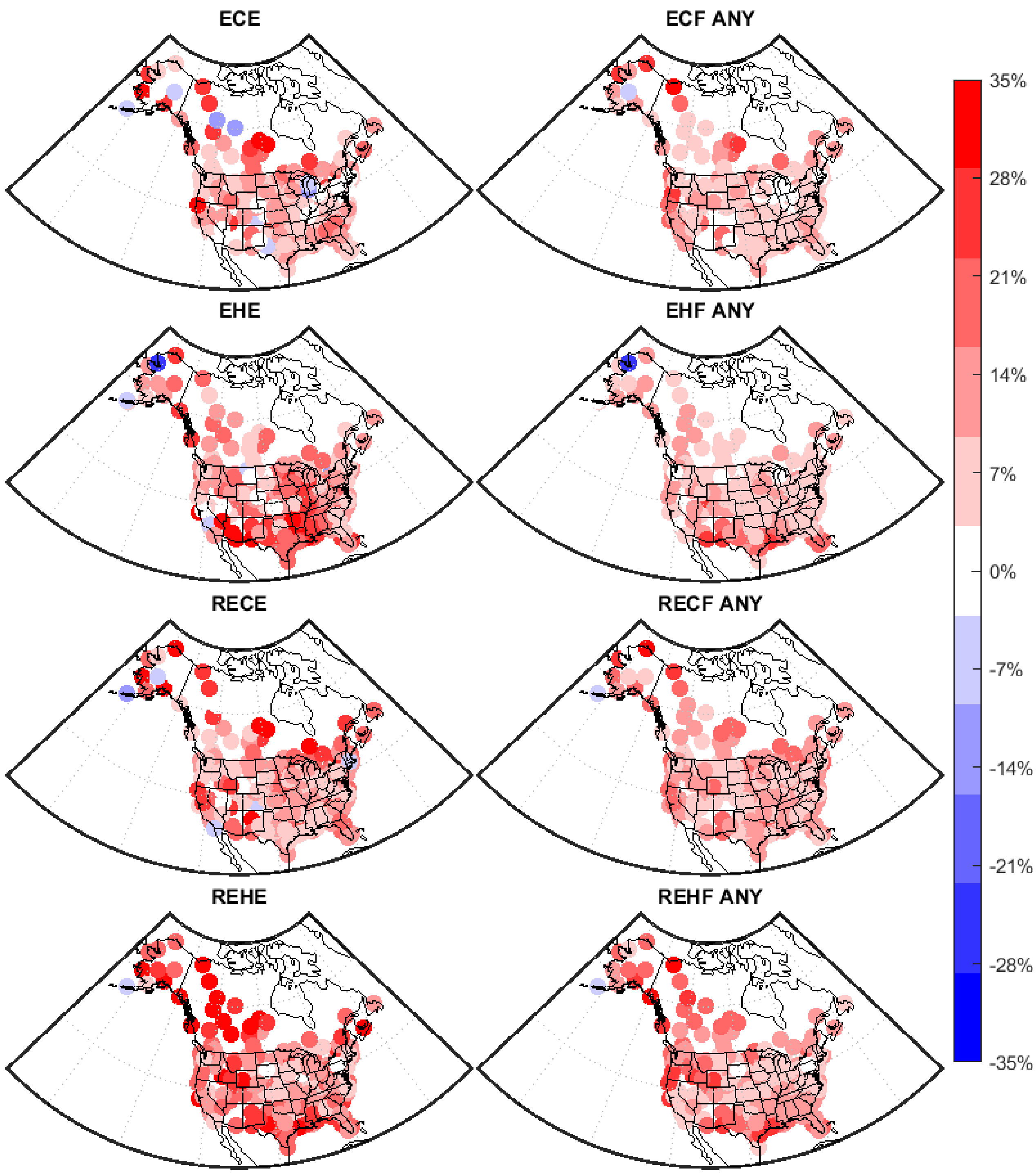

Figure S1. Like Figure 1, except for difference in match percentage points between ERA5 and NARR. Red (blue) colors show ERA5 is better (worse) than NARR at matching station-defined events.

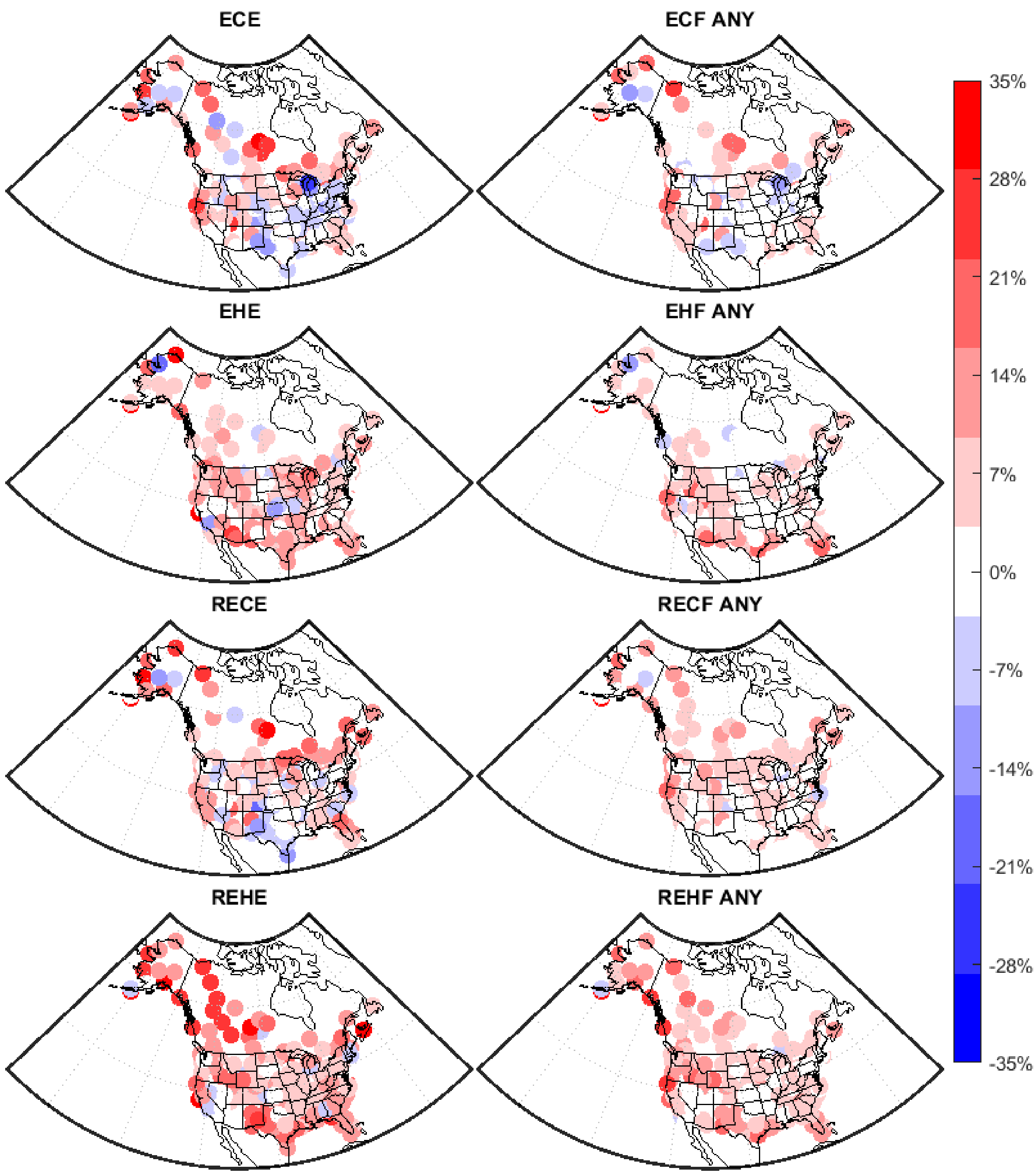

Figure S2. Like Figure 1, except for difference in match percentage points between ERA5-LAND and NARR. Red (blue) colors show ERA5-LAND is better (worse) than NARR at matching station-defined events.

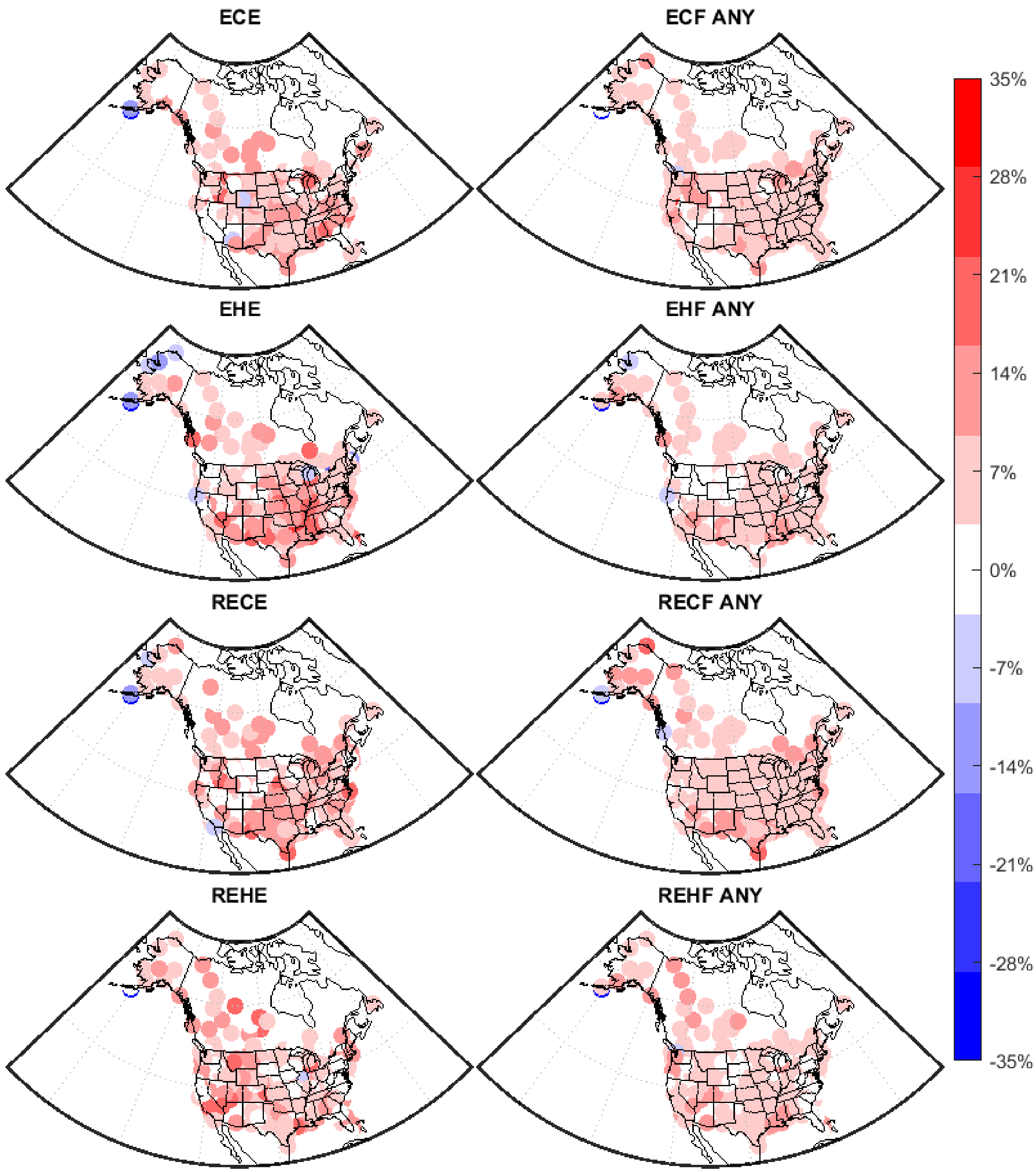

Figure S3. Like Figure 1, except for difference in match percentage points between ERA5 and ERA5-LAND. Red (blue) colors show ERA5 is better (worse) than ERA5-LAND at matching station-defined events.

Table 1. Median and range of annual frequencies for all data sets, and overall match percentage representing the percent of days identified using the station-based data that were also identified by each reanalysis data set.

Event	Station-based			NARR			ERA5			ERA5-LAND			Match percentage with station		
	Median	Range		Median	Median		Median	Median		Median	Range		NARR	ERA5	ERA5-LAND
ECE	2.16	1.53	2.71	2.33	2.33	2.33	2.25	1.61	2.63	2.33	1.83	2.61	72.0%	81.4%	74.0%
ECF > 0	14.43	11.37	17.03	15.56	15.56	15.56	14.95	11.92	17.42	15.56	12.28	17.33	81.3%	88.8%	82.5%
EHE	2.39	1.95	3.08	2.47	2.47	2.47	2.39	2	3.08	2.47	2.11	3.03	57.5%	72.2%	63.9%
EHF > 0	15.87	12.97	20.74	16.42	16.42	16.42	15.92	13.37	20.53	16.42	14.11	20.25	74.2%	82.7%	77.1%
RECE	2.3	1.47	2.89	2.47	2.47	2.47	2.37	1.47	3	2.47	1.72	3.06	66.9%	78.6%	70.5%
RECE > 0	15.16	10.63	19.03	16.42	16.42	16.42	15.76	10.87	20.11	16.42	11.44	20.33	71.5%	82.2%	74.8%
REHE	2.58	1.97	4.39	2.75	2.75	2.75	2.64	2.08	3.84	2.75	2.14	3.67	61.0%	76.5%	69.6%
REHF > 0	17.5	13.34	28.76	18.36	18.36	18.36	17.63	13.89	25.61	18.36	14.31	24.36	66.8%	78.7%	72.8%