PROCEEDINGS A

rspa.royalsocietypublishing.org

Research

Article submitted to journal

Subject Areas:
Climatology, machine learning

Keywords:
machine learning, climate change, drought

Author for correspondence:
Insert corresponding author name
e-mail: kate.marvel@nasa.gov

Using Machine Learning to Identify Novel Hydroclimate States

Kate Marvel[1,2] and Benjamin I. Cook[2]

[1]Columbia University Center for Climate Systems Research
[2]NASA Goddard Institute for Space Studies

Anthropogenic emissions of greenhouse gases, ozone depleting substances, and aerosols as well as changes to land cover and land use are expected to alter drought risk in the future. However, droughts are not uncommon or unprecedented, as documented in tree-ring based reconstructions of the summer average Palmer Drought Severity Index (PDSI). Using an unsupervised machine learning method trained on these reconstructions of pre-industrial climate, we identify outliers: years in which the spatial pattern of PDSI is unusual relative to "normal" variability. We show that in many regions, outliers are more frequently identified in the twentieth and twenty-first centuries. This trend is more pronounced when the regional drought atlases are combined into a single global dataset. By definition, outlier patterns at the 10% level are expected to occur once per decade, but from 1950-2000 more than six years per decade are identified as outliers in the global drought atlas. Extending the global drought atlas through 2020 using an observational dataset suggests that anomalous global drought conditions are present in 80% of years in the 21st century. Our results indicate, without recourse to climate models, that the world is more frequently experiencing drought conditions that are highly unusual in the context of past natural climate variability.

THE ROYAL SOCIETY
PUBLISHING

1. Introduction

Anthropogenic climate change is expected to substantially affect the risk of drought across the globe [1,2]. A large body of literature has identified the “fingerprints” of anthropogenic influence- particularly emissions of aerosols and greenhouse gases- on multiple aspects of the hydrological cycle, including increases in atmospheric water vapor [3], shifts in the intensity and pattern of global rainfall [4–6], increases in extreme precipitation [7], reductions in snow cover [8,9], changes in the seasonal cycle of precipitation [10] and evaporation [11,12], and drought risk on both global [13,14] and regional scales [15]. The credibility of these detection and attribution studies, however, largely rests on the accuracy and credibility of the internal and natural variability generated by general circulation models used to differentiate forced signals from noise [16].

There are alternate methods for assessing pre-industrial hydroclimate variability. High-quality, tree-ring-based drought atlases reconstruct Common Era hydroclimate conditions over a large portion of the global land area [17–25]. These paleoclimate datasets target the Palmer Drought Severity Index (PDSI), a widely used metric of soil moisture drought, and provide an assessment of naturally forced and internal drought variability that extends back for centuries. These long-term records offer opportunities to place current trends in a long-term context and provide a means for differentiating forced signals from pre-industrial variability without the use of general circulation model (GCM) output.

New widely available statistical and software tools can help us to extract useful information from these invaluable datasets. The methods and techniques of machine learning have found many applications in climate science, from improving model parameterizations [26] to observationally constraining various quantities [27] to identifying features such as tropical cyclones [28] and atmospheric rivers [29] in data. Here, we use a simple anomaly detection algorithm, trained on tree ring-based drought reconstructions, to identify outlier years in which drought conditions are unusual compared to pre-industrial variability. This method identifies multivariate outliers: unusual spatial patterns, rather than spatial averages. The goal of this work is not to identify why certain years are outliers, or to attribute outlying datapoints to specific modes of internal variability or external forcing. Such attribution requires the use of additional models, whether simple theoretical representations or the output of complex general circulation models (GCMs), to generate the (unobservable) responses to external forcing or pure internal variability, and will be the subject of further work. Instead, we construct a multivariate detection method that does not rely on GCM estimates of preindustrial or naturally forced variability to establish the baseline of “normal” climate conditions.

2. Methods

We use an “Isolation Forest” [30] to detect outliers in drought atlases. This widely used machine learning algorithm exploits the properties of outliers: that they are, by definition, different from “normal” states, and that they are few in number. In an Isolation Forest, a randomly selected line, plane, or (in multiple dimensions) hyperplane divides multi-dimensional data into two branches; this dividing process is then repeated until a point (or line, or feature) is isolated. It takes more cuts- more branches on the random decision tree- to isolate typical points, whereas outliers will be quickly isolated. There are several advantages to this method. First, it is unsupervised, meaning that we do not need to explicitly identify outliers in the training dataset. Second, the only user decision is the intuitive and easily specified offset parameter, which specifies the fraction of observations that should be designated as outliers in the training dataset. Finally, it is fast and simple to generalize to high-dimensional data. We compared our results to an alternate machine learning method (graphical lasso with cross-validation) that explicitly learns the spatial covariance of the training dataset, as well as a one class support vector machine for unsupervised anomaly detection, and find that it delivers similar results in much faster time.

To train and validate the algorithm, we use tree ring-based drought atlases that reconstruct PDSI over Europe and the Mediterranean (the Old World Drought Atlas; OWDA [21]), European Russia (ERDA [24]), North America (NADA [19,20]), Mexico (MXDA) [25], monsoon Asia (MADA) [17], eastern Australia and New Zealand (ANZDA [18]), and South America (SADA) [23]. PDSI is averaged over the summer season (JJA in the northern hemisphere, DJF in the southern); in the southern hemisphere, the year corresponds to the calendar year of the respective January. Each atlas has been regridded to a common 2.5° latitude-longitude grid using an interpolation that preserves the area-weighted mean and truncated, if necessary, to span the years 1400-2005. We also create a "global" drought atlas (GDA) by combining data from each regional drought atlas. In cases where there is spatial overlap, the more recent MXDA, which includes more proxies over the domain, supersedes the NADA, and the more recent ERDA supersedes the OWDA.

Drought alases provide reliable, millennial-scale records of past variability and have been extensively validated against known historical droughts and pluvials (e.g. [17]). However, they do have limitations: the quality of reconstructions is not always uniform across the spatial domain, and their target variable, PDSI, excludes possible plant responses to increased carbon dioxide [31]. Additionally, it is possible that higher carbon dioxide concentrations may influence drought atlas data through direct influences on trees, either by increasing water use efficiency or greening in respnse to elevated CO_2. There is considerable uncertainty in the degree to which tree growth is decoupled from drought in a high-CO_2 world, but previous model-based [12,32,33] and observational [34–36] studies show that plant water use responses often enhance surface drying, even when CO_2 effects on water use efficiency are accounted for.

We divide each drought atlas into a training period 1400-1849 and a recent period 1850-2000. The isolation forest learns over the initial training period, in which we have 450 years of spatial patterns for each drought atlas. We grow a decision tree by recursively partitioning the training dataset until each year is isolated at one of the nodes. Anomalous years are easily isolated while more typical years require longer path lengths along the decision tree. To formalize the problem, we neglect the temporal covariance structure of the data and consider the training dataset (which spans n_t time steps and n_s grid cells) as simply n_t instances of an n_s-variate distribution. The anomaly score of a particular map $x(t)$ at time t is defined as

$$s(x(t), n_t) = -2^{-<h(x)t))>/c(n_t)}$$

where $< h(x) >$ is the average number of branches required to isolate the node corresponding to the spatial pattern $x(t)$ and $c(n_t)$ is the average path length of an unsuccessful search. If the spatial pattern at time t is a strong outlier, then very few branches are required to isolate it and $s(x(t), n_t)$ will be close to -1. By contrast, if $x(t)$ is a typical sample, then it will take many cuts to isolate it, $< h(x) >$ will approach the maximum and $s(x(t), n_t)$ will approach zero. The decision function D to classify inliers and outliers depends on an offset parameter ν:

$$D(x, n) = s(x, n) - \nu \tag{2.1}$$

which is defined in order to obtain the expected number of outliers (samples with decision function < 0) in the training dataset.

The training period of the Isolation Forest serves only to set the threshold beyond which a map or other data point is considered to be an outlier. For example, if we set the contamination parameter ν to be 0.01, this sets the threshold for the decision function such that 1% of years over the training period are considered to be outliers. To extend drought atlas data to 2020, we use instrumental PDSI values from the CRU TS dataset [37,38]. While drought atlases are calibrated using recent instrumental data, the two datasets are not identical, and there are some differences in the decision functions between the instrumental record and the reconstructions. However, in all cases a year identified as an outlier in one dataset is also flagged as an outlier in another, albeit occasionally at a different level.

This analysis begs the question: what, in practice, does it mean for a datapoint to be an outlier? The degree to which PDSI patterns are "unusual" depends on what is considered "usual": the range of conditions to which local infrastructure, policy, and strategy is adapted. This depends on many things: geography, population, and political boundaries, for example. Some impacts of a severe agricultural drought in one small region may be offset by infrastructure connections to adjacent regions or trade links with farther-away places. Conversely, highly unusual conditions—wet or dry– that occur simultaneously across multiple regions may stress existing social, physical, and economic structures developed under particular assumptions about the limits of climate variability. For example, Figure 1(a) shows years identified as outliers at the 1% level for the average PDSI in the Mediterranean region over the 1400-1849 training period (all regions are defined using the IPCC WG 1 definitions available at `https://github.com/IPCC-WG1/Atlas`), all of which correspond to unusually wet or dry years in the region. Figure 1(b) shows years identified as outliers at the 1% level in the simplified two-dimensional spatial domain consisting of the Mediterranean and Northern Europe. The years of most extreme wetness or dryness in the Mediterranean are classified as outliers, but so too are years of relatively normal conditions in the Mediterranean, if those years were characterized by anomalous conditions in northern Europe. Figure Figure 1(c) shows years identified as outliers at the 1% level in the simplified spatial domain consisting of the Mediterranean, northern Europe, and western-central Europe. For the Mediterranean region, years designated as outliers at the 1% level are the wet

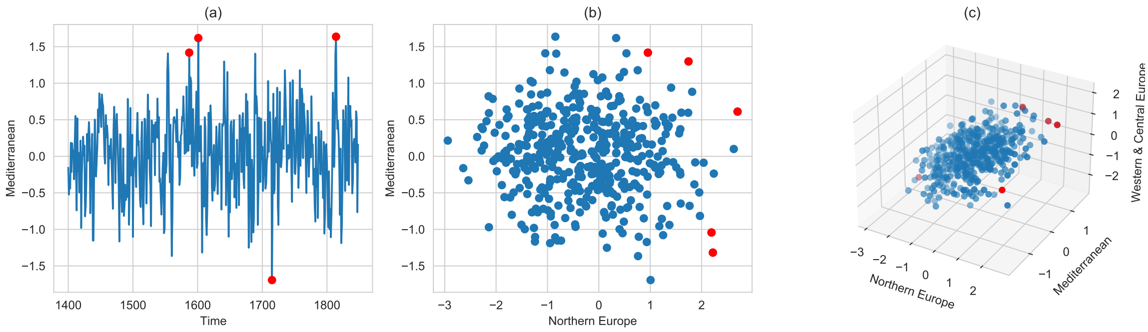

Figure 1. (a): Outliers at the 10% level for regional average PDSI in the Mediterranean. (b) Outliers at the 10% level in the simplified 2-dimensional spatial domain consisting of the average PDSI over Northern Europe and the average PDSI over the Mediterranean region (c) Outliers at the 10% level for the simplified 3D spatial domain consisting of average PDSI over Northern Europe, the Mediterranean, and western-central Europe.

years 1587, 1601, and 1715, and the dry year 1814. For the two-dimensional region, the 1% outlier years are 1587, 1607, 1641, 1686, and 1734. For the three-dimensional region, they are 1419, 1528, 1607, 1686, and 1734. These differences reflect the shifting definition of "outlier" as the spatial dimension of the region expands as well as the flexibility of the isolation forest method. Given a region of importance, the method quickly isolates unusual patterns: hydroclimate states that stand out from the variability of the "normal" climate of that region. The notion of an outlier is specific to the scale and characteristics of the region considered.

3. Results

(a) North America

To illustrate the method and results for high-dimensional data, we first focus on the North American drought atlas (NADA, [19,20]). Figure 2(a) shows the regional average summer (JJA) PDSI, along with years flagged as outliers at the 1% level, while Figure 2(b) shows the decision function used to classify years as outliers or inliers. The definition of outlier depends on the parameter ν (Equation 2.1). A year is classified as an outlier if the decision function is less than zero; the zero line is chosen such that a user-determined fraction of years in the training period are outliers. Outliers identified in the training period include the years 1635, 1670,1718, 1736, and 1756. While some of these years (e.g. 1756) are associated with extrema in the regional mean, others are associated with unusual smaller-scale patterns. For example, while conditions in the western portion of the continent resembled a typical El Nino response [22,39] in 1635, the year was characterized by severe droughts in modern-day Alaska, central-western Canada, the upper Midwest, and the northeast. In the recent period, the largest outlier identified in the NADA is 1934 (Figure 2(d): a known heat wave and drought event associated with the height of the Dust Bowl [40]; the "El Nino of the century" 1997-8 is also identified as an outlier. To extend the NADA

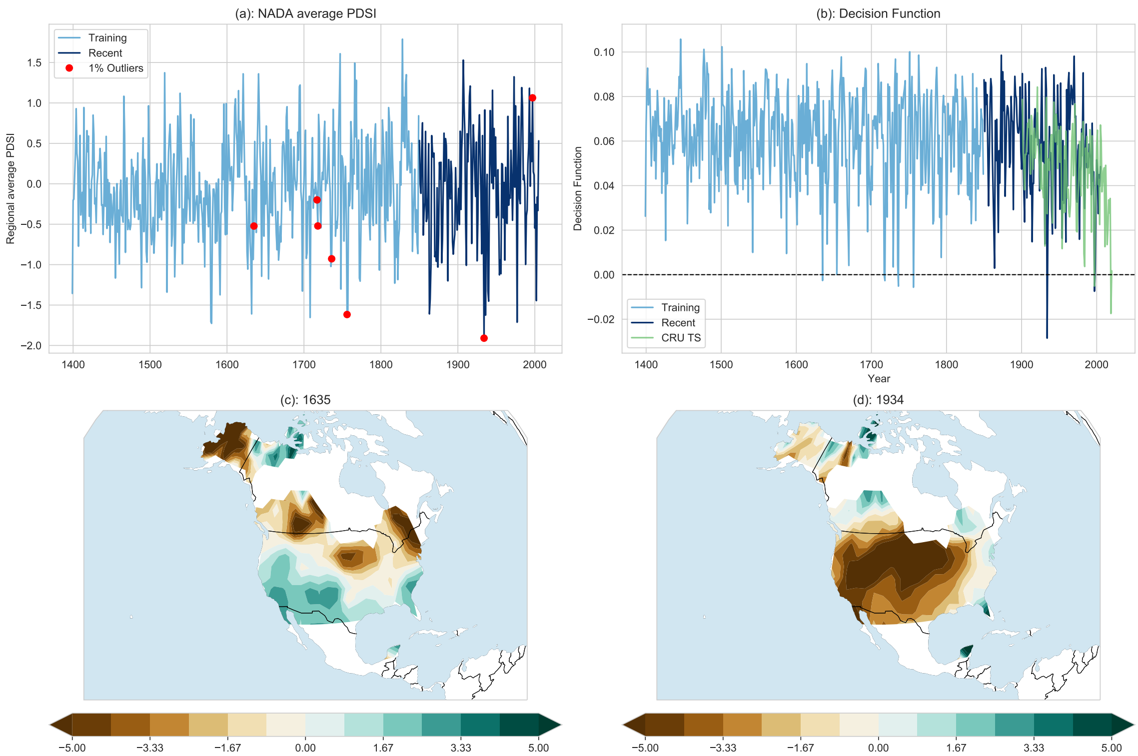

Figure 2. (a): NADA regional mean Palmer drought severity index (PDSI) over the training period (light blue) and recent period (dark blue). Years identified as outliers at the 1% level are shown in red. (b) Decision function calculated by the isolation forest, where the zero line is determined by an offset chosen by requiring 1% of the years in the training dataset (light blue) to be outliers. Also shown are the decision function of 1901-2000 drought atlas data (dark blue) and CRU TS data over the NADA region extended to 2020 (green). (c) An outlier obtained from the training dataset (d) An outlier in the recent drought atlas dataset.

reconstructions to the present day, we apply the isolation forest algorithm to patterns in the CRU TS PDSI dataset derived from meteorological observations. Because the NADA and the CRU dataset are not identical over the time periods in which they overlap, we first test whether the CRU data is more likely to be classified as an outlier by an algorithm trained on tree-ring-derived reconstructions. This does not appear to be the case; in the time period in which the datasets overlap, fewer years (1984 and 1997) are identified as outliers in CRU TS than in NADA. Twenty-first century CRU data show a steep decrease in the decision function, with 2019 and 2020 both flagged as outliers. This indicates a trend in recent decades toward more unusual North American drought conditions.

(b) Mexico and the southern United States

In the Mexican Drought Atlas [25], the years identified as anomalous in the training period are 1455, 1668 (widespread drought, part of a longer drought in the 1660s for which there is much historical documentation [20]), 1746 (wet anomalies in the north, particularly in Baja California, 1816 (extremely wet conditions throughout most of the region, and 1819 (identified as the most anomalous year by the isolation forest, Figure 3). Only one year (1956) is identified as an outlier in the recent period, and neither CRU TS nor the drought atlas data show a drastic shift toward more anomalous conditions. Unlike the NADA, with which it partially overlaps, the MXDA covers a region in which climate variability is relatively spatially coherent, and therefore outlier years represent anomalously wet or dry conditions across the entire region rather than more complex spatial patterns. This is apparent in Figure 3a, which shows that all identified outlier years in this region correspond to extrema in the regional mean PDSI.

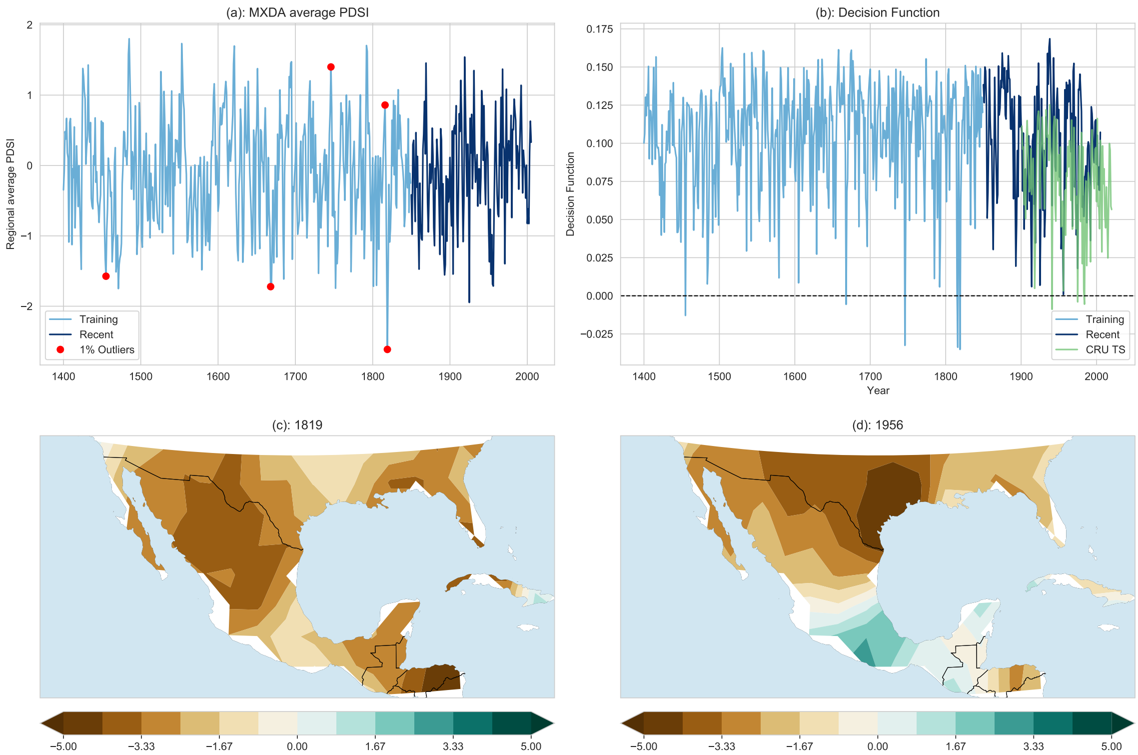

Figure 3. (As in Figure 1, but for the Mexican Drought Atlas (MXDA)

(c) Europe and the Mediterranean

There is a similar trend toward more unusual conditions in the region covered by the Old World Drought Atlas ([21], OWDA). In the training period for this dataset, the years 1503 (characterized by severe drought in Central Europe as documented in [41]), 1601 (wet conditions in southern Europe and northeastern Scandivavia), 1607 (drought throughout much of the Ottoman empire and wet conditions elsewhere) 1660 (severe drought in Anatolia), and 1686 (severe drought in southeastern Europe and extreme wetness in Scandinavia) are identified as outliers at the 1% level, with 1686 the largest outlier. The relative concentration of outliers in the seventeenth century speaks to the importance of anomalous conditions associated with the Little Ice Age during this time. Interestingly, none of these years overlap with the subset of historically recorded

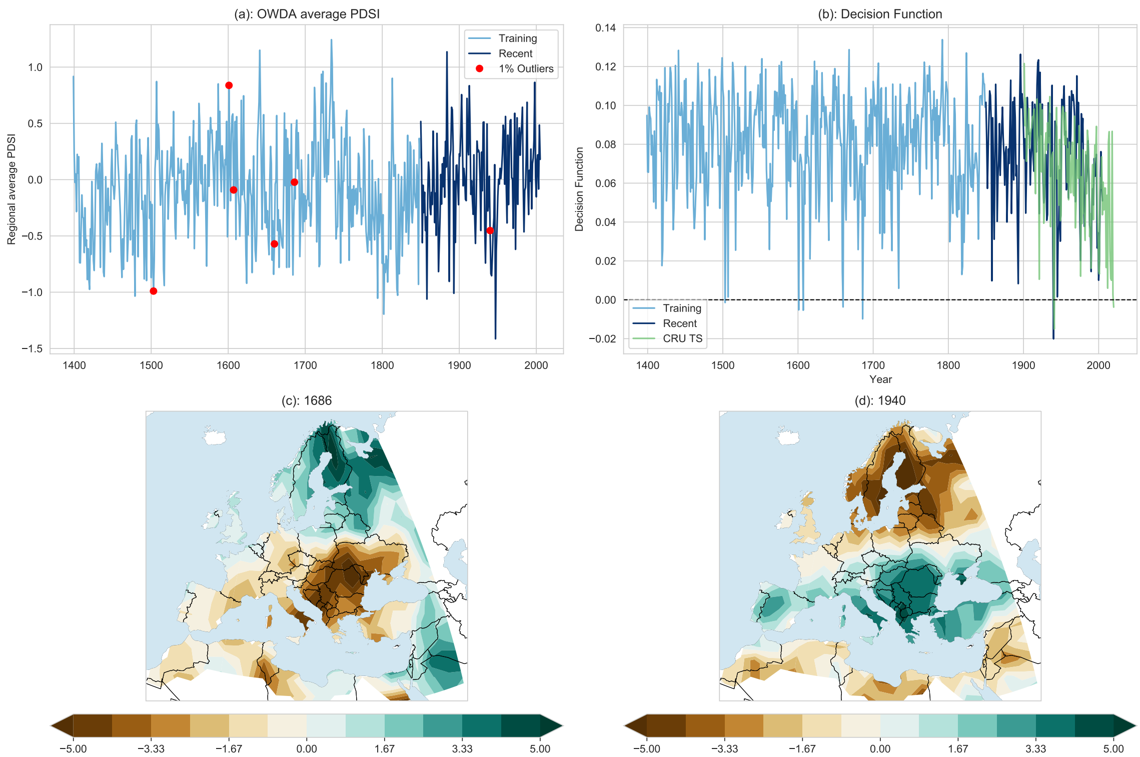

Figure 4. As in Figure 1, but for the Old World Drought Atlas (OWDA)

extreme hydroclimate events considered in [21] (Cook et al Sci Adv), although the Great Drought of 1893 and the Czech Lands drought of 1616 considered in that work are both identified as outliers by the isolation forest at the 5% threshold and the 1921 drought in Central Europe recently discussed in [42] is identified as an outlier at the 10% threshold. This does not mean those events were not extreme or that their impacts were not severe, merely that they were more regional in extent, and that the corresponding patterns over the entire OWDA domain were not particularly anomalous. For the 1850-2005 period, only one year (1940) in the OWDA is identified as an outlier. Over the 1850-2020 period, three years (1940, 1941, and 2020) in the CRU TS dataset are identified as outliers. However, the decision function for the CRU TS dataset over the OWDA region, as with the NADA region, shows a decline in the twenty-first century.

(d) European Russia

The European Russia drought atlas, bounded by the East European Plain and the Ural mountains, extends and updates the Old World Drought Atlas. [24] use rotated EOF analysis to identify three main spatial modes of variability across this domain: interior European Russia (VF1), Finland and the Baltics (VF2), and the southwestern region of the ERDA domain (VF3). They identify extreme dry and wet years associated with each pattern. Similarly, the isolation forest identifies 1408 (the driest year in VF2), 1453 (driest year in VF3), 1466 (wettest year in VF1), 1772 (wettest year in

VF3), and 1802 (3rd driest year in VF1) as 1% outliers over the 1400-1849 training period. Over the recent period, the years 1891, 1921, 1936, 1939, 1941, and 1975 are identified as outliers. This is consistent with [24], who find that each of these years corresponds to an extreme dry or wet year in one of the VF domains and identified the 1921 pan-ERDA drought and the 1936 drought in eastern European Russia as strong outliers. By contrast with the OWDA and NADA, however, the recent CRU TS dataset does not show a marked trend toward more unusual conditions in this region.

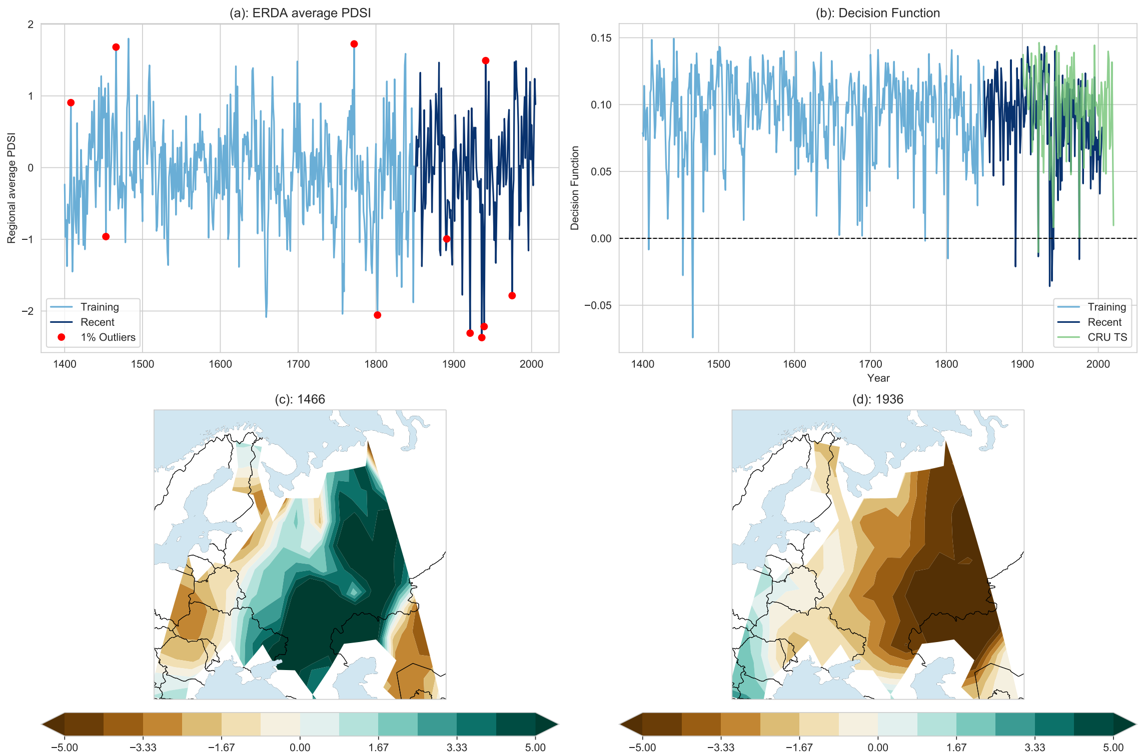

Figure 5. As in Figure 1, but for the European Russia Drought Atlas (ERDA)

(e) Monsoon Asia

Unlike the European Russia case, Figure 6(a) shows that outliers are strongly clustered toward the end of the record for the region covered by the Monsoon Asia drought atlas (MADA). None of the historical events considered in [17] - the Ming Dynasty drought (1638 to 1641), the Strange Parallels drought (1756 to 1768), the East India drought (1790 to 1796) , and the late Victorian Great Drought (1876 to 1878)- are flagged as outliers at the 1% or 5% level *in the wider region*, although 1790 is identified as an outlier at the 10% level. This does not imply that these extreme events were not impactful– all of them are well-documented in the historical record and apparent in the reconstructions. However, it does imply that these events were not associated with particularly unusual or unprecedented spatial patterns of PDSI across the entire MADA domain. The years in the training dataset identified as 1% outliers by the isolation forest are 1430 and 1705 (characterized by wet anomalies in northeastern China and Mongolia), 1495 (the most anomalous year, Figure 6 c), and 1626 and 1785 (severe drought in the westernmost portion of the region and wet anomalies in northern and central India and southeast Asia). Far more years are identified as 1% outliers in the recent dataset: 1917, 1956, 1957, 1958, 1959, 1960, 1981, 1987, 1990, 1991, 1992, 1993, 1994, 1998, 2001, 2002, and 2003. CRU TS data show a similar drastic shift toward more anomalous conditions over the twentieth century, with an extremely steep drop in the decision function beginning in the 1950s.

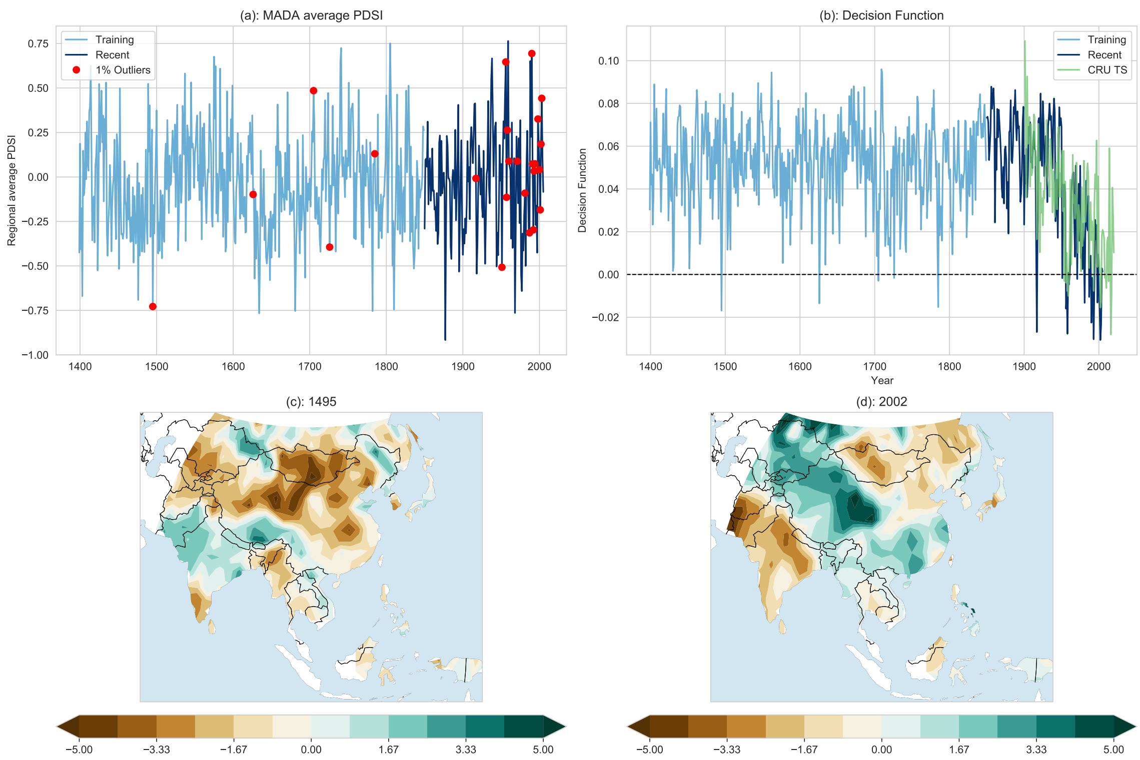

Figure 6. As in Figure 1, but for the Monsoon Asia Drought Atlas (MADA)

(f) Australia and New Zealand

In the southern hemisphere, the drought atlases reconstruct austral summer (DJF) average PDSI. For the Australia and New Zealand drought atlas (ANZDA), the isolation forest identifies 1405, 1412, 1573, 1743, and 1833 as outliers; all but 1833 are associated with wet anomalies, not drought (Figure 7a). Similarly, all outlier years in the recent 1850-2005 period are not associated with drought conditions: 1956, 1971, 1974, and 1976 were all extremely wet years. The fact that wet years are far more likely than dry years to be identified as outliers is commensurate with previous findings that the unusually wet period from the 1950s through the 1970s has few analogues in the extended drought atlas reconstruction [18]. While the differences between CRU and drought atlas data are substantial in this region (Figure 7b), both show a trend toward more unusual conditions over the twentieth and twenty-first centuries.

(g) South America

For the South American drought atlas (SADA), two years identified as outliers at the 1% level are associated with regional mean wet conditions: 1498 (the most unusual year, Figure 8c) and 1594, while the outlier years 1595 , 1622, and 1696 are all associated with dry regional mean anomalies (although neither the driest or wettest years in the regional average are identified as outliers. The

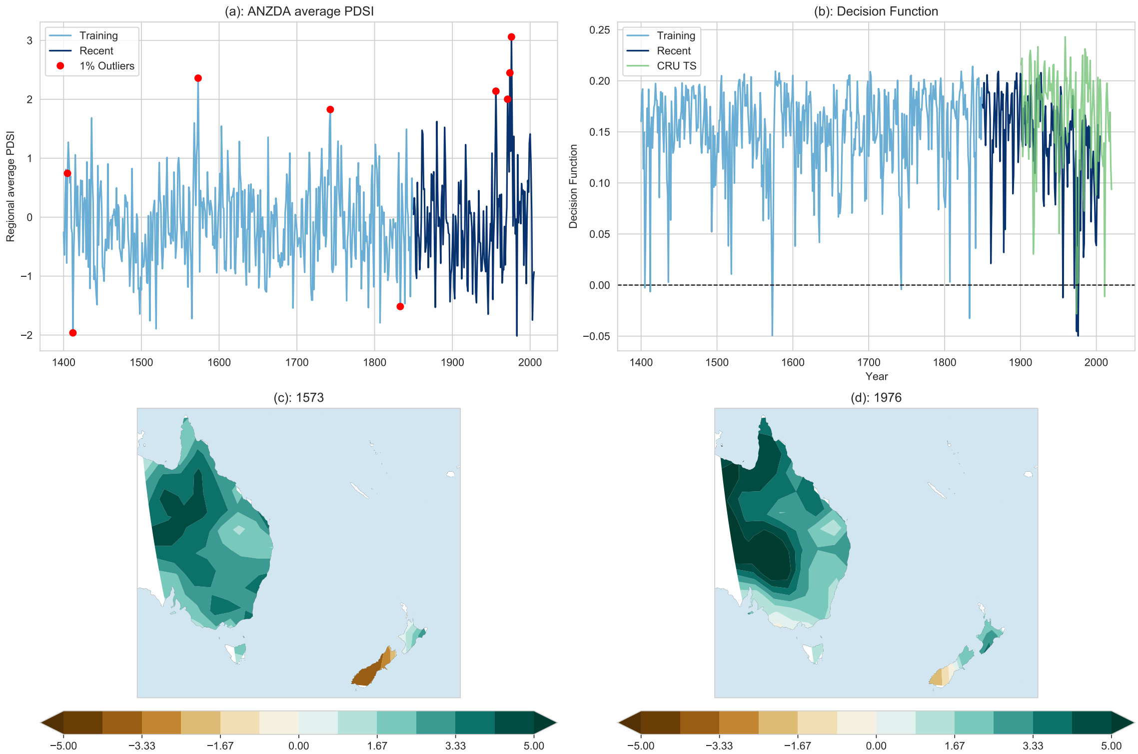

Figure 7. As in Figure 1, but for the Australia and New Zealand drought atlas (ANZDA)

years 1967 and 1968 are identified as the only outliers in the recent period, and are associated with a well-documented drought in the region [23].

(h) Global

The global drought atlas, obtained by regridding and combining all publicly available drought atlases as in [43], shows a marked shift to more unusual conditions over the twentieth and twenty-first centuries. While the (quasi) global mean PDSI shows large oscillations over the historical time period, far more outliers are detected in recent decades, many of which are not associated with global mean extremes (Figure 9a). The decision function decreases rapidly over the twentieth century (Figure 9b) The most unusual year in the recent drought atlas data is 1998, which at the time was the strongest El Nino event on record. This year is also identified as an outlier in the CRU TS data- as are 2001-3, 2010, 2011, 2015-17, 2019, and 2020. The strongest outlier in the 1901-2020 CRU TS dataset is 2016 (Figure 9 d) The binary classification of years as outliers or inliers is somewhat arbitrary by definition, and depends on the specification of the offset parameter that determines the fraction of outliers in the training dataset. Moreover, there is a certain degree of stochasticity in the isolation forest algorithm, meaning that a pattern that barely exceeds the outlier threshold and is classified as an outlier in one iteration may be classified as an inlier in another. However, clear patterns emerge from the global drought atlas, regardless of the threshold for outliers. Figure 10a shows the number of years in each decade flagged as outliers at the 10% level. In the 15th and 16th centuries, most decades contain 0 to 2 outlier years. There is a distinct increase in the number of 10% outliers in the seventeenth century, and a sustained increase in these outliers throughout the twentieth century. A similar pattern holds for outliers at the 1% level (Figure 10b). Notably, in the most recent decade, six years are identified as strong outliers. While the global drought atlas lacks data from Africa and western Australia and is therefore

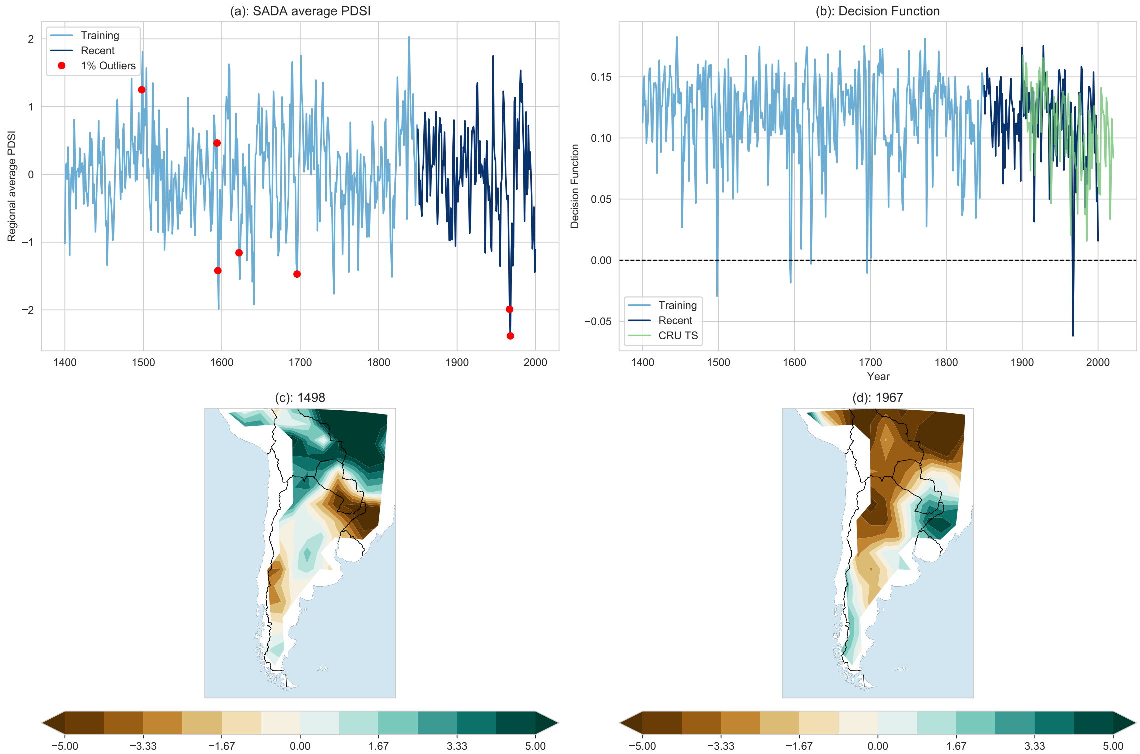

Figure 8. As in Figure 1, but for the South American drought atlas (SADA)

biased toward the northern hemisphere, the trend toward more unusual conditions is observable in both hemispheres. Figure 11 shows the decision function for JJA-average PDSI in the northern hemisphere (blue) and for DJF-average PDSI in the southern hemisphere (red). Because there is less data in the southern hemisphere, the decision function is somewhat noisier, and thus we require a larger offset to classify a set percentage of years in the training period as outliers. However, both the northern hemisphere and the southern hemisphere drought atlases show clear declines over the twentieth century. Interestingly, after sharply decreasing throughout the 1950s and 1960s, the southern hemisphere decision function appears to reach a minimum in 1975. The trend in the decision function over the monsoon Asia drought atlas is steeply decreasing and is extremely strong compared to the other drought atlases. To test the sensitivity of our results to the inclusion of MADA, we compile a version of the Global Drought Atlas that excludes this data. Figure 12 shows the decision function for the drought atlas and CRU TS data excluding and including MADA. While the decrease in the decision function is less steep in the global drought atlas when MADA is excluded, the CRU TS dataset shows nearly identical trends when MADA is included or excluded. We therefore conclude that despite the rapid, early, and pronounced emergence of novel conditions over monsoon Asia, this region is not the primary driver of global anomalies. Instead, global PDSI patterns appear increasingly anomalous in the context of pre-industrial variability.

(i) Robustness to training period

Farther back in time, the number of tree ring chronologies decreases and the reconstructions become more sparse. Additionally, changes in the quality, quantity and spatial coverage of the underpinning meteorological data may affect our results. When we identify recent years as outliers, we must be careful to specify the period and conditions relative to which they are

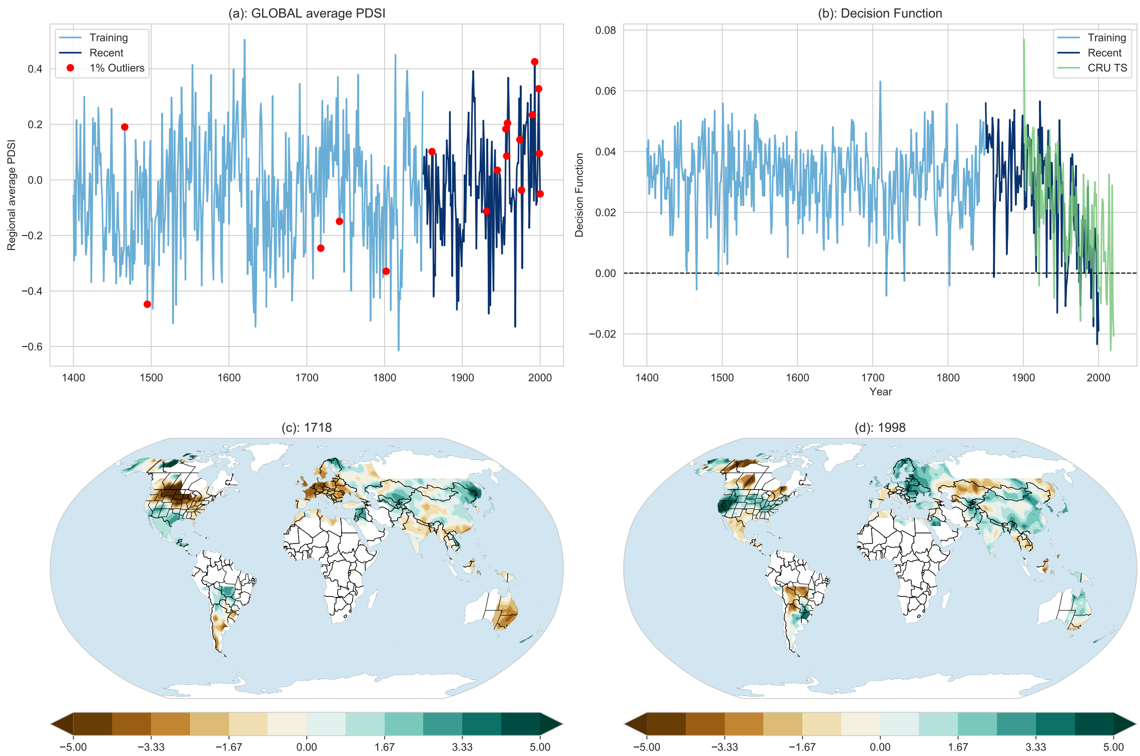

Figure 9. As in Figure 1, but for the global drought atlas (GDA)

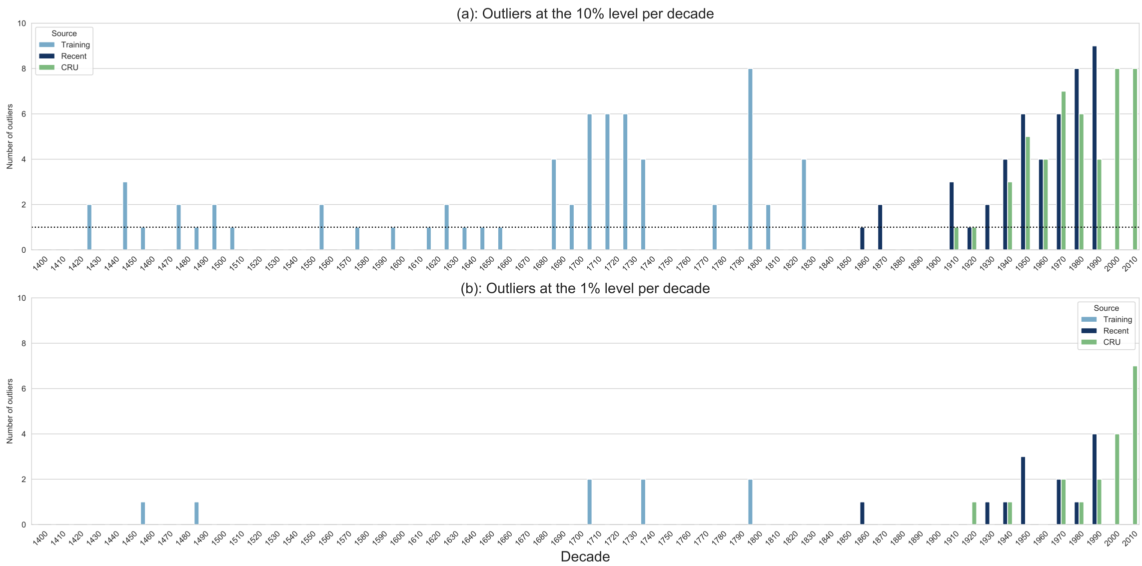

Figure 10. Number of years per decade identified as outliers at the (a) 10% and (b) 1% levels

unusual. If inhomogeneities in the data are present, an identified outlier might be an artifact of more complete recent data compared to less complete, accurate, or homogenous past data. To

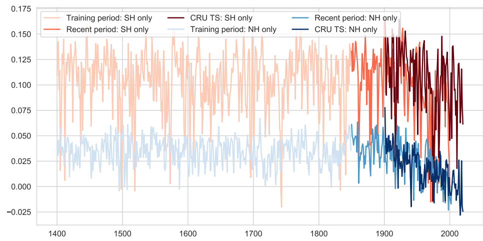

Figure 11. Decision functions for drought atlas data over the northern (blue) and southern (red) hemispheres

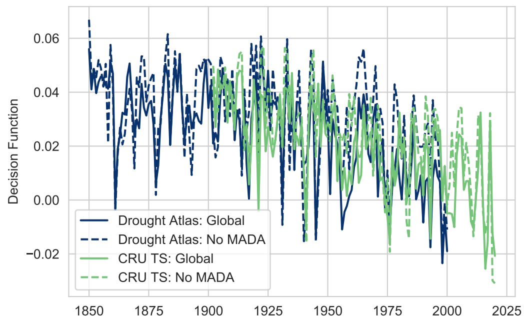

Figure 12. Decision functions for the global drought atlas including and excluding the Monsoon Asia drought atlas

test this, we perform sensitivity analyses targeting a training period of 1900-1950, a period of mostly homogenous data coverage for the global drought atlas. We note that unlike the pre-1850 reconstructions, this training period does not reflect the sole influence of preindustrial (naturally forced and internal) variability: greenhouse gases and aerosol concentrations were both increasing over this period. Figure 13 shows the anomaly scores $s(x,n)$ for the global drought atlas trained

Figure 13. Anomaly scores $s(x,n)$ for the global drought atlas trained over the 1400-1849 period (orange) and the 1900-1950 period (blue). The decision function is defined as $D(x,n) = s(x,n) - \nu$. The offsets ν are chosen such that $D(x,n)$ is less than zero for 1% of years in the training period.

over the 1400-1849 period (orange) and the shorter, more recent period (blue). The thresholds ν(dotted lines) are chosen such that the decision function (Eq. 2.1) is negative for an outlier at the 1% level. The 1900-1950 training period contains not only fewer years, but occurs coincident with increasing external forcing: this affects the baseline definition of 'usual", and the single year identified as an outlier (1917) relative to this period is not identified as an outlier relative to the longer 1400-1849 period. However, regardless of the training period chosen, there is a trend toward lower scores, and thus recent years are more likely to be identified as outliers relative both to preindustrial variability and to conditions in the first half of the twentieth century.

Finally, we train the algorithm on CRU TS meterological data (masked to have the same coverage as the global drought atlas) spanning the latter half of the twentieth century (1950-1999). Over these fifty years, the five years identified as outliers at the 10% level are 1960, 1962, 1972, 1975, and 1976. When applied to 21 years out-of-sample data, fully ten years are identified as outliers: 2001, 2002, 2003, 2008, 2009, 2010, 2011, 2015, 2016, and 2020. Evidently, twenty-first century conditions as reflected in the CRU TS meteorological dataset are notably unusual even in comparison with recent preceding decades.

4. Conclusion

Detection of a trend in outlier frequency, for whatever reason, does not necessarily imply detection of anthropogenic influence or natural external forcing. The isolation forest algorithm is not designed to identify specific characteristics of patterns identified as anomalous, nor is it particularly useful for attribution research. Regional and global hydroclimate responses to greenhouse gases, aerosols, ozone depletion, and natural forcings have been identified across multiple variables and at multiple scales. Because the sustained emergence of novel conditions coincides with increases in anthropogenic forcing, it is reasonable to suspect a human fingerprint on these conditions, and indeed, multiple studies have suggested a strong link between human activities and shifts in hydroclimate patterns [13–15,43,44]. The temporal structure of detected anomalies also suggest possible roles for multiple anthropogenic forcings. For example, ozone depletion is known to have affected Southern hemisphere rainfall by forcing changes in the atmospheric circulation [45–48], while greenhouse gas-forced warming increases global precipitation and evaporation [4,5,49] and anthropogenic aerosols directly and indirectly change the surface energy balance, affecting both moisture availability and transport [50–52]. Greenhouse gas emissions have steadily increased, while the 1991 Montreal Protocol largely halted ozone depletion. Meanwhile, aerosol emissions have decreased globally and shifted from North America and western Europe to south and east Asia [53]. Additional human activities such as changes in land use and cover have likely also affected PDSI in many regions [54–56]. In particular, the rapid shift to anomalous conditions in the MADA in the 1950s may be associated with large-scale irrigation in that region [57].

However, attribution requires an a priori knowledge of the response pattern to a particular external forcing or collection of forcings- something that is not observable, given that the real-world climate is a combination of forced response and internal variability. Such fingerprints are generally obtained using climate model simulations, and therefore the credibility of attribution studies rests on the credibility of these models. Additionally, attribution is complicated by interactions between forcing agents: if, for example, aerosol forcing induces dynamical shifts that modify the circulation on which other forcings act, or if greenhouse gas-induced warming affects the surface energy balance response to aerosols or irrigation, the question of attribution to a single forcing is ill-posed. The machine learning methods presented here can supplement and support existing detection and attribution studies by detecting outliers and novel conditions in a model-independent way. Our methods capitalize on the existence of long drought atlas datasets that provide a wealth of spatiotemporal information throughout the Common Era, allowing an algorithm trained on pre-industrial variability to learn, in an unsupervised manner and with minimal methodological choices, what "normal" patterns of hydroclimate variability might look like. Future work will further advance our ability to place current trends in long-term historical context, integrating the paleoclimate record, model simulations, and recent observations to not only detect emergent properties of the climate system, but more robustly attribute them as well.

Our results indicate a major shift in global hydroclimate in recent decades. Years which do not appear to be particularly anomalous in any specific region appear as strong outliers on hemispheric or global scales, reinforcing previous conclusions that while regional signals have not necessarily emerged from the noise, in the larger picture, the world is already experiencing more unusual drought conditions. While the results obtained using the isolation forest are consistent with previous literature that uses more traditional methods [43], they yield information about the time at which novel conditions emerge but not about the conditions themselves. A more complete spatial understanding of these conditions is required in order to identify the underlying physical drivers of the anomalous patterns. Simply put, we not only wish to *detect* novelty in the data, but *characterize* it as well. Future work will build on these methods to learn about the spatial patterns driving the emergence of novel conditions, complementing existing detection and attribution methods and bolstering confidence in the drivers of a changing hydroclimate.

Data Accessibility. The regional drought atlases used in this study are available from the Tree-Ring Drought Atlas Portal at `http://drought.memphis.edu`. The most recently updated versions of ANZDA (`https://www.dropbox.com/s/nrizk1a1a289awh/anzdaV2.nc`) and MADA (`https://www.dropbox.com/s/n2lo99h9qn17prg/madaV2.nc`) that are used in this analysis are available from the indicated websites. CRU TS data is available for download at `https://catalogue.ceda.ac.uk/uuid/c26a65020a5e4b80b20018f148556681`.

Authors' Contributions. K.M. designed the study, carried out the analysis, and drafted the manuscript. B.I.C. provided a critical review of the manuscript, updated drought atlas data, and substantial intellectual contributions to the understanding of drought atlas outliers.

Competing Interests. The authors declare that they have no competing interests.

Funding. The authors were supported by the NOAA MAPP grant, "Integrating models, paleoclimate, and recent observations to develop process-level understanding of projected changes in U.S. drought." This work benefited from the authors' participation in the NOAA CMIP6 Task Force.

Acknowledgements. We are grateful to Sonali McDermid and Celine Bonfils for useful discussions. We thank the NOAA National Centers for Environmental Information for making the paleo data accessible and the National Science Foundation Paleoclimatology Program, Paleo Perspectives on Climate Change for supporting the development of the Tree-Ring Drought Atlas Portal.

References

1. Cook BI, Mankin JS, Marvel K, Williams AP, Smerdon JE, Anchukaitis KJ. 2020 Twenty-first century drought projections in the CMIP6 forcing scenarios. *Earth's Future* **8**, e2019EF001461.
2. Cook BI, Mankin JS, Anchukaitis KJ. 2018 Climate Change and Drought: From Past to Future. *Current Climate Change Reports* **4**, 164–179.
3. Santer BD, Mears C, Wentz F, Taylor K, Gleckler P, Wigley T, Barnett T, Boyle J, Brüggemann W, Gillett N et al.. 2007 Identification of human-induced changes in atmospheric moisture content. *Proceedings of the National Academy of Sciences* **104**, 15248–15253.
4. Zhang X, Zwiers FW, Hegerl GC, Lambert FH, Gillett NP, Solomon S, Stott PA, Nozawa T. 2007 Detection of human influence on twentieth-century precipitation trends. *Nature* **448**, 461–465.
5. Marvel K, Bonfils C. 2013 Identifying external influences on global precipitation. *Proceedings of the National Academy of Sciences* **110**, 19301–19306.
6. Schurer AP, Ballinger AP, Friedman AR, Hegerl GC. 2020 Human influence strengthens the contrast between tropical wet and dry regions. *Environmental Research Letters* **15**, 104026.
7. Donat MG, Lowry AL, Alexander LV, OGorman PA, Maher N. 2016 More extreme precipitation in the worlds dry and wet regions. *Nature Climate Change* **6**, 508.
8. Dettinger MD, Cayan DR. 1995 Large-scale atmospheric forcing of recent trends toward early snowmelt runoff in California. *Journal of Climate* **8**, 606–623.
9. Barnett TP, Adam JC, Lettenmaier DP. 2005 Potential impacts of a warming climate on water availability in snow-dominated regions. *Nature* **438**, 303–309.
10. Marvel K, Biasutti M, Bonfils C, Taylor KE, Kushnir Y, Cook BI. 2017 Observed and Projected Changes to the Precipitation Annual Cycle. *Journal of Climate* **30**, 4983–4995.

11. Marvel K, Cook BI, Bonfils C, Smerdon JE, Williams AP, Liu H. 2021 Projected Changes to Hydroclimate Seasonality in the Continental United States. *Earth's Future* **9**, e2021EF002019.
12. Mankin JS, Smerdon JE, Cook BI, Williams AP, Seager R. 2017 The Curious Case of Projected Twenty-First-Century Drying but Greening in the American West. *Journal of Climate* **30**, 8689–8710.
13. Bonfils CJ, Santer BD, Fyfe JC, Marvel K, Phillips TJ, Zimmerman SR. 2020 Human influence on joint changes in temperature, rainfall and continental aridity. *Nature Climate Change* **10**, 726–731.
14. Chiang F, Mazdiyasni O, AghaKouchak A. 2021 Evidence of anthropogenic impacts on global drought frequency, duration, and intensity. *Nature communications* **12**, 1–10.
15. Williams AP, Cook ER, Smerdon JE, Cook BI, Abatzoglou JT, Bolles K, Baek SH, Badger AM, Livneh B. 2020 Large contribution from anthropogenic warming to an emerging North American megadrought. *Science* **368**, 314–318.
16. Santer BD, Mears C, Doutriaux C, Caldwell P, Gleckler P, Wigley T, Solomon S, Gillett N, Ivanova D, Karl T et al.. 2011 Separating signal and noise in atmospheric temperature changes: The importance of timescale. *Journal of Geophysical Research: Atmospheres (1984–2012)* **116**.
17. Cook ER, Anchukaitis KJ, Buckley BM, DArrigo RD, Jacoby GC, Wright WE. 2010 Asian monsoon failure and megadrought during the last millennium. *Science* **328**, 486–489.
18. Palmer JG, Cook ER, Turney CS, Allen K, Fenwick P, Cook BI, ODonnell A, Lough J, Grierson P, Baker P. 2015 Drought variability in the eastern Australia and New Zealand summer drought atlas (ANZDA, CE 1500–2012) modulated by the Interdecadal Pacific Oscillation. *Environmental Research Letters* **10**, 124002.
19. Cook ER, Seager R, Heim RR, Vose RS, Herweijer C, Woodhouse C. 2010 Megadroughts in North America: Placing IPCC projections of hydroclimatic change in a long-term palaeoclimate context. *Journal of Quaternary Science* **25**, 48–61.
20. Cook ER, Seager R, Cane MA, Stahle DW. 2007 North American drought: Reconstructions, causes, and consequences. *Earth-Science Reviews* **81**, 93–134.
21. Cook ER, Seager R, Kushnir Y, Briffa KR, Büntgen U, Frank D, Krusic PJ, Tegel W, van der Schrier G, Andreu-Hayles L et al.. 2015 Old World megadroughts and pluvials during the Common Era. *Science advances* **1**, e1500561.
22. Baek SH, Smerdon JE, Coats S, Williams AP, Cook BI, Cook ER, Seager R. 2017 Precipitation, temperature, and teleconnection signals across the combined North American, Monsoon Asia, and Old World Drought Atlases. *Journal of Climate* **30**, 7141–7155.
23. Morales MS, Cook ER, Barichivich J, Christie DA, Villalba R, LeQuesne C, Srur AM, Ferrero ME, González-Reyes Á, Couvreux F et al.. 2020 Six hundred years of South American tree rings reveal an increase in severe hydroclimatic events since mid-20th century. *Proceedings of the National Academy of Sciences* **117**, 16816–16823.
24. Cook ER, Solomina O, Matskovsky V, Cook BI, Agafonov L, Berdnikova A, Dolgova E, Karpukhin A, Knysh N, Kulakova M et al.. 2020 The european russia drought atlas (1400–2016 ce). *Climate Dynamics* **54**, 2317–2335.
25. Stahle DW, Cook ER, Burnette DJ, Villanueva J, Cerano J, Burns JN, Griffin D, Cook BI, Acuña R, Torbenson MC et al.. 2016 The Mexican Drought Atlas: Tree-ring reconstructions of the soil moisture balance during the late pre-Hispanic, colonial, and modern eras. *Quaternary Science Reviews* **149**, 34–60.
26. Gentine P, Pritchard M, Rasp S, Reinaudi G, Yacalis G. 2018 Could machine learning break the convection parameterization deadlock?. *Geophysical Research Letters* **45**, 5742–5751.
27. Ceppi P, Nowack P. 2021 Observational evidence that cloud feedback amplifies global warming. *Proceedings of the National Academy of Sciences* **118**.
28. Chen R, Zhang W, Wang X. 2020 Machine learning in tropical cyclone forecast modeling: A review. *Atmosphere* **11**, 676.
29. Muszynski G, Kashinath K, Kurlin V, Wehner M et al.. 2019 Topological data analysis and machine learning for recognizing atmospheric river patterns in large climate datasets. *Geoscientific Model Development* **12**, 613–628.
30. Liu FT, Ting KM, Zhou ZH. 2008 Isolation forest. In *2008 eighth ieee international conference on data mining* pp. 413–422. IEEE.
31. Swann AL, Hoffman FM, Koven CD, Randerson JT. 2016 Plant responses to increasing CO2 reduce estimates of climate impacts on drought severity. *Proceedings of the National Academy of Sciences* **113**, 10019–10024.

32. Mankin JS, Seager R, Smerdon JE, Cook BI, Williams AP, Horton RM. 2018 Blue water trade-offs with vegetation in a CO2-enriched climate. *Geophysical Research Letters* **45**, 3115–3125.
33. Mankin JS, Seager R, Smerdon JE, Cook BI, Williams AP. 2019 Mid-latitude freshwater availability reduced by projected vegetation responses to climate change. *Nature Geoscience* **12**, 983–988.
34. Deng Y, Wang S, Bai X, Luo G, Wu L, Chen F, Wang J, Li C, Yang Y, Hu Z et al.. 2020 Vegetation greening intensified soil drying in some semi-arid and arid areas of the world. *Agricultural and Forest Meteorology* **292**, 108103.
35. Trancoso R, Larsen JR, McVicar TR, Phinn SR, McAlpine CA. 2017 CO 2 -vegetation feedbacks and other climate changes implicated in reducing base flow. *Geophysical Research Letters* **44**, 2310–2318.
36. Ukkola AM, Prentice IC, Keenan TF, Van Dijk AI, Viney NR, Myneni RB, Bi J. 2016 Reduced streamflow in water-stressed climates consistent with CO2 effects on vegetation. *Nature Climate Change* **6**, 75–78.
37. van der Schrier G, Barichivich J, Briffa KR, Jones PD. 2013 A scPDSI-based global data set of dry and wet spells for 1901-2009. *Journal of Geophysical Research: Atmospheres* **118**, 4025–4048.
38. Dunn RJH, Aldred F, Gobron N, Miller JB, Willett KM, Ades M, Adler R, Richard PA, Allan R, Anderson J, Argüez A, Arosio C, Augustine JA, Azorin-Molina C, Barichivich J, Beck HE, Becker A, Bellouin N, Benedetti A, Berry DI, Blenkinsop S, Bock O, Bodin X, Bosilovich MG, Boucher O, Buehler SA, Calmettes B, Carrea L, Castia L, Christiansen HH, Christy JR, Chung ES, Coldewey-Egbers M, Cooper OR, Cornes RC, Covey C, Cretaux JF, Crotwell M, Davis SM, de Jeu RAM, Degenstein D, Delaloye R, Girolamo LD, Donat MG, Dorigo WA, Durre I, Dutton GS, Duveiller G, Elkins JW, Fioletov VE, Flemming J, Foster MJ, Frith SM, Froidevaux L, Garforth J, Gentry M, Gupta SK, Hahn S, Haimberger L, Hall BD, Harris I, Hemming DL, Hirschi M, pen (Ben) Ho S, Hrbacek F, Hubert D, Hurst DF, Inness A, Isaksen K, John VO, Jones PD, Junod R, Kaiser JW, Kaufmann V, Kellerer-Pirklbauer A, Kent EC, Kidd R, Kim H, Kipling Z, Koppa A, Kraemer BM, Kratz DP, Lan X, Lantz KO, Lavers D, Loeb NG, Loyola D, Madelon R, Mayer M, McCabe MF, McVicar TR, Mears CA, Merchant CJ, Miralles DG, Moesinger L, Montzka SA, Morice C, Mösinger L, Mühle J, Nicolas JP, Noetzli J, Noll B, O'Keefe J, Osborn TJ, Park T, Pasik AJ, Pellet C, Pelto MS, Perkins-Kirkpatrick SE, Petron G, Phillips C, Po-Chedley S, Polvani L, Preimesberger W, Rains DG, Randel WJ, Rayner NA, Rémy S, Ricciardulli L, Richardson AD, Robinson DA, Rodell M, Rodríguez-Fernández NJ, Rosenlof K, Roth C, Rozanov A, Rutishäuser T, Sánchez-Lugo A, Sawaengphokhai P, Scanlon T, Schenzinger V, Schlegel RW, Sharma S, Shi L, Simmons AJ, Siso C, Smith SL, Soden BJ, Sofieva V, Sparks TH, Stackhouse PW, Steinbrecht W, Stengel M, Streletskiy DA, Sun-Mack S, Tans P, Thackeray SJ, Thibert E, Tokuda D, Tourpali K, Tye MR, van der A R, van der Schalie R, van der Schrier G, van der Vliet M, van der Werf GR, Vance A, Vernier JP, Vimont IJ, Vömel H, Vose RS, Wang R, Weber M, Wiese D, Wilber AC, Wild JD, Wong T, Woolway RI, Zhou X, Yin X, Zhao G, Zhao L, Ziemke JR, Ziese M, Zotta RM. 2021 Global Climate. *Bulletin of the American Meteorological Society* **102**, S11–S142.
39. St. George S, Meko DM, Cook ER. 2010 The seasonality of precipitation signals embedded within the North American Drought Atlas. *The Holocene* **20**, 983–988.
40. Cook BI, Seager R, Smerdon JE. 2014 The worst North American drought year of the last millennium: 1934. *Geophysical Research Letters* **41**, 7298–7305.
41. Ionita M, Dima M, Nagavciuc V, Scholz P, Lohmann G. 2021 Past megadroughts in central Europe were longer, more severe and less warm than modern droughts. *Communications Earth & Environment* **2**, 1–9.
42. van der Schrier G, Allan RP, Ossó A, Sousa PM, Van de Vyver H, Van Schaeybroeck B, Coscarelli R, Pasqua AA, Petrucci O, Curley M et al.. 2021 The 1921 European drought: impacts, reconstruction and drivers. *Climate of the Past* **17**, 2201–2221.
43. Marvel K, Cook BI, Bonfils CJ, Durack PJ, Smerdon JE, Williams AP. 2019 Twentieth-century hydroclimate changes consistent with human influence. *Nature* **569**, 59–65.
44. Seager R, Hoerling M, Schubert S, Wang H, Lyon B, Kumar A, Nakamura J, Henderson N. 2015 Causes of the 2011 to 2014 California drought. *Journal of Climate*.
45. Kang S, Polvani L, Fyfe J, Sigmond M. 2011 Impact of polar ozone depletion on subtropical precipitation. *Science* **332**, 951–954.
46. Min SK, Son SW. 2013 Multi-model attribution of the Southern Hemisphere Hadley cell widening: Major role of ozone depletion. *Journal of Geophysical Research: Atmospheres*.

47. Polvani LM, Waugh DW, Correa GJ, Son SW. 2011 Stratospheric ozone depletion: The main driver of twentieth-century atmospheric circulation changes in the Southern Hemisphere. *Journal of Climate* **24**, 795–812.
48. Shindell DT, Schmidt GA. 2004 Southern Hemisphere climate response to ozone changes and greenhouse gas increases. *Geophysical Research Letters* **31**.
49. Durack PJ, Wijffels SE, Matear RJ. 2012 Ocean salinities reveal strong global water cycle intensification during 1950 to 2000. *Science* **336**, 455–458.
50. Bollasina MA, Ming Y, Ramaswamy V. 2011 Anthropogenic aerosols and the weakening of the South Asian summer monsoon. *science* p. 1204994.
51. Polson D, Bollasina M, Hegerl G, Wilcox L. 2014 Decreased monsoon precipitation in the Northern Hemisphere due to anthropogenic aerosols. *Geophysical Research Letters* **41**, 6023–6029.
52. Wilcox LJ, Highwood EJ, Dunstone NJ. 2013 The influence of anthropogenic aerosol on multi-decadal variations of historical global climate. *Environmental Research Letters* **8**, 024033.
53. Turnock S, Allen R, Andrews M, Bauer S, Emmons L, Horowitz L et al.. 2020 Historical and future changes in air pollutants from CMIP6 models. *Atmospheric Chemistry and Physics Discussions* pp. 1–40.
54. Al-Yaari A, Ducharne A, Thiery W, Cheruy F, Lawrence D. 2021 The role of irrigation expansion on historical climate change during the last 115 years: insights from CMIP6. .
55. Cook BI, Shukla SP, Puma MJ, Nazarenko LS. 2015 Irrigation as an historical climate forcing. *Climate Dynamics* **44**, 1715–1730.
56. Singh D, McDermid SP, Cook BI, Puma MJ, Nazarenko L, Kelley M. 2018 Distinct Influences of Land Cover and Land Management on Seasonal Climate. *Journal of Geophysical Research: Atmospheres* **123**, 12,017–12,039.
57. Ambika AK, Wardlow B, Mishra V. 2016 Remotely sensed high resolution irrigated area mapping in India for 2000 to 2015. *Scientific data* **3**, 1–14.