On the optimal design of field significance tests for changes in climate extremes

Jianyu Wang, Chao Li, Francis W. Zwiers, Xuebin Zhang, Guilong Li, Zhihong Jiang, Panmao Zhai, Ying Sun, Zhen Li & Qun Yue

2021

Original citation:

Wang, J., Li, C., Zwiers, F. W., Zhang, X., Li, G., Jiang, Z., Zhai, P., Sun, Y., Li, Z., & Yue, Q. (2021). On the Optimal Design of Field Significance Tests for Changes in Climate Extremes. *Geophysical Research Letters, 48*(9). https://doi.org/10.1029/2021GL092831

AGU ADVANCING EARTH AND SPACE SCIENCE

Geophysical Research Letters

RESEARCH LETTER
10.1029/2021GL092831

On the Optimal Design of Field Significance Tests for Changes in Climate Extremes

Jianyu Wang[1,2], Chao Li[1,2,4], Francis Zwiers[3,4], Xuebin Zhang[5], Guilong Li[5], Zhihong Jiang[4], Panmao Zhai[6], Ying Sun[7], Zhen Li[8], and Qun Yue[1,2]

[1]Key Laboratory of Geographic Information Science, Ministry of Education, East China Normal University, Shanghai, China, [2]School Geographic Sciences, East China Normal University, Shanghai, China, [3]Pacific Climate Impacts Consortium, University of Victoria, Victoria, BC, Canada, [4]Nanjing University of Information Science and Technology, Nanjing, China, [5]Climate Research Division, Environment and Climate Change Canada, Toronto, ON, Canada, [6]State Key Laboratory of Severe Weather, Chinese Academy of Meteorological Sciences, China Meteorological Administration, Beijing, China, [7]Laboratory for Climate Studies, National Climate Center, China Meteorological Administration, Beijing, China, [8]Key Laboratory of Regional Climate-Environment for Temperate East Asia, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China

Key Points:

- Field significance is generally determined by summarizing the results of individual tests conducted at different locations in a domain
- Inconsistent field significance conclusions can be drawn when using different local test methods
- For extreme precipitation, field significance determined from the simple Mann-Kendall test performs better than other commonly used ones

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

C. Li,
cli@geo.ecnu.edu.cn

Citation:

Wang, J., Li, C., Zwiers, F., Zhang, X., Li, G., Jiang, Z., et al. (2021). On the optimal design of field significance tests for changes in climate extremes. *Geophysical Research Letters*, *48*, e2021GL092831. https://doi.org/10.1029/2021GL092831

Received 3 FEB 2021
Accepted 21 APR 2021

Abstract Field significance tests have been widely used to detect climate change. In most cases, a local test is used to identify significant changes at individual locations, which is then followed by a field significance test that considers the number of locations in a region with locally significant changes. The choice of local test can affect the result, potentially leading to conflicting assessments of the impact of climate change on a region. We demonstrate that when considering changes in the annual extremes of daily precipitation, the simple Mann-Kendall trend test is preferred as the local test over more complex likelihood ratio tests that compare the fits of stationary and nonstationary generalized extreme value distributions. This lesson allows us to report, with enhanced confidence, that the intensification of annual extremes of daily precipitation in China since 1961 became field significant much earlier than previously reported.

Plain Language Summary Changes to weather and climate extremes at individual locations can be highly uncertain due to natural variability. Much of the natural variability in precipitation extremes occurs on small spatial scales, and thus analyzing changes at different locations in a region with a field significance test can help extract information about changes in the region that is less affected by natural variability. An important component of doing so is the local test that is used to identify significant changes at individual locations and a field significance test that evaluates whether such changes are found at more locations than would be expected from natural variability in an unchanged climate. By contrasting several common local test methods with varying complexity, we find that the simple Mann-Kendall test tends to yield a field significance test with high power of detection. Based on these lessons, we find that the intensification of extreme precipitation in China became field significant much earlier than previously reported, thereby resolving uncertainty about whether intensification is in fact discernable in China.

1. Introduction

Field significance (Livezey & Chen, 1983) and false discovery rate (Wilks, 2006) tests are widely used to detect climate change (e.g., Alexander et al., 2006; Chen et al., 2021; E. M. Fischer & Knutti, 2014; Kiktev et al., 2003; Li, Jiang, et al., 2018; Lorenz et al., 2019; Risser et al., 2019; Sun et al., 2020; Westra et al., 2013; W. Zhang & Zhou, 2019). We focus on field significance tests here because they have been used in recent studies of climate change in China and because they continue to be used heavily. The motivation for their use is that while changes to a climate variable at individual stations can be highly uncertain due to unforced internal climate variability, statistically significant changes at individual stations may nevertheless occur more frequently across a large domain than would be consistent with a stationary climate where all variability is the result of natural, unforced, internal processes, that is, internal climate variability.

The method involves a null hypothesis that the climate is stationary. Under this hypothesis, any observed trends that are statistically significant are considered as random manifestations of unforced internal climate variability. Thus, field significance is assessed by using the fraction of stations or grid boxes showing

significant trends in a region as a test statistic. A spatiotemporal block bootstrap procedure is implemented to infer the null distribution of this test statistic. Specifically, a large ensemble of samples of plausible observations under internal variability is generated via bootstrapping, which reshuffles the observations by time such that the long-term trends in the observations are removed while their spatial and some aspects of temporal dependence are retained (e.g., von Storch & Zwiers, 1999; see supporting information for details). The local assessments of trends in the observed and bootstrapped samples can be conducted in different ways. When considering changes in extremes over time, these might involve a trend detection method such as the Mann-Kendall test for an overall tendency (e.g., Alexander et al., 2006; Kiktev et al., 2003; Wang & Swail, 2001; W. Zhang & Zhou, 2019; X. Zhang & Zwiers, 2004) or the likelihood ratio test examining the fit of a nonstationary extreme value distribution to these extremes (e.g., Chen et al., 2021; Li et al., 2018; Sun et al., 2020; Westra et al., 2013; Zhang et al., 2010). If the observed fraction of stations with significant trends turns out to be unusual in the distribution of the bootstrapped fractions, that is, smaller than a given field significance level, the observed trends are claimed to be field significant.

The choice of local trend detection method can affect the final field significance evaluation, since different tests may operate at different local significance levels and may be more or less able to detect changes. That is, different local tests may have different *size* and *power*. The size of a test is the probability of falsely rejecting the null hypothesis. Note that the actual size of the test, i.e., its operating significance level, may differ from the significance level that was specified. The power, which is affected by the size, is the probability of correctly rejecting the null hypothesis when it is false.

How to optimally conduct a field significance test so as to improve the power of field detection remains unclear. In recent years, field significance tests have been increasingly used for detecting changes in climate extremes at regional and even smaller scales (e.g., Chen et al., 2021; W. Li & Chen 2020; Li et al., 2018; Lorenz et al., 2019; Sun et al., 2020; Zhang et al., 2020; W. Zhang & Zhou, 2019). Internal variability is a large component of the variation of most climate extremes at small spatial scales (e.g., E. M. Fischer et al., 2014; Li et al., 2019, 2021), thus challenging the detection of their changes. A field significance detection method that efficiently synthesizes results from individual locations is therefore of great value.

Taking the detection of changes in precipitation extremes in China as an example, we consider the sensitivity of the assessment of field significance to the choice of local detection method, and show that the simple Mann-Kendall test is a relatively favorable choice after contrasting the size and power of field significance detection based on some other common local detection methods. We also clarify that although increasing significance levels of local tests can improve the power of local detection, doing so does not strongly affect the power of field detection. Based on these lessons, we revisit whether precipitation extremes in China, as represented by annual maxima of daily precipitation (Rx1day), have intensified significantly over time based the historical records up to 2017. Our analysis of annual values of Rx1day has not explicitly considered the impact of variations in the seasonality of precipitation extremes in China. Although we focus on a particular variable in a particular region, the reported results have general implications to other climate variables in different regions.

2. Data and Methods

2.1. Data

We acquired daily precipitation observations at 839 National Reference and Basic Stations in mainland China for the period 1961–2017 from the National Meteorological Information Center of China. We excluded stations with more than 30 missing values in any 1 year during the whole period, leaving 574 stations for this study, for which the missing value rate is less than 1% in almost all years (Figure S1). The retained observations were homogenized following the method used in Li et al. (2015) before extracting Rx1day values for each station.

We also use Rx1day from a large ensemble of 50 initial-condition simulations by the Canadian Earth System Model version 5 (CanESM5) with historical forcings (Swart et al., 2019) to study the size and power of field significance tests constructed with different local tests. In this case we consider the slightly shorter period 1961–2014 as this is the last year of the available historical simulations.

3. Methods

The overall procedure for field significance detection is as presented at the beginning of the letter. Here we briefly describe the three common local trend detection methods that we consider; more details can be found in the supporting information.

We consider the nonparametric Mann-Kendall (MK) test (Kendall, 1975; Mann, 1945), which asks whether there is a general increasing or decreasing tendency in the data to which it is applied. An assumption concerning the distribution of the data is not required.

We also consider two methods that stem from statistical extreme value theory (e.g., R. A. Fischer & Tippett, 1928; Leadbetter et al., 1983), which suggests that the distributions of block maxima will converge to the generalized extreme value (GEV) distribution *as blocks become large if convergence occurs*. The GEV distribution is therefore often used to approximate the distribution of annual maxima, such as Rx1day. Thus, another approach for testing for trend in a timeseries of block maxima is to use a likelihood ratio (LR) test (e.g., Coles, 2001) to determine whether a GEV distribution with time-varying parameters provides a significantly better fit to the data than a GEV distribution with constant parameters. We consider a GEV distribution with a location parameter that is a linear function of time and a distribution that also allows its scale parameter to vary with time in a log-linear form. In the following, we refer to LR tests with these two distributions as LR1 and LR2, respectively. Although the extended GEV distribution for the r-largest maxima in a block, $r > 1$, has also been used to detect changes in precipitation extremes (e.g., Li et al., 2018), consideration of single block maxima is adequate for clarifying our ideas.

Physical laws suggest that climate warming should intensify extreme precipitation over most landmasses (e.g., Allen & Soden, 2008). We therefore focus on the detection of increases in Rx1day. The LR test does not distinguish between positive and negative trends, and thus it is a two-sided test, in contrast to the MK test, which can be performed as one-sided test in which rejection occurs when the observed trend is greater than a specified positive critical value. In order to ensure a fair comparison, we conduct a local two-sided LR test to identify whether there is a significant trend at a given local significance level a, and, if so, to determine if the trend is positive. Doing so, the effective nominal local significance level is $a/2$, given that positive and negative trends are equally likely under the null hypothesis that only unforced internal climate variability is playing a role.

To evaluate the size and power of a field significance test built with a local detection method, we use the 50-member initial-condition ensemble of historical CanESM5 simulations to construct 5,000 realizations of the Rx1day field over China for the period 1961–2014 influenced by internal variability alone and a further 5,000 realizations that are also influenced by the historical external forcings (see supporting information for details). The size and power of a field significance test can be estimated as the rejection rates of the test when applied to these two sets of constructed simulations, respectively. Simultaneously, the size and power of the local tests can be estimated at grid cells.

As an attractive test, its estimated size should be consistent with the specified significance level and its power should be high. The interpretation of test results relies on the assumption that in the absence of a climate change signal, rejections occur at a rate that corresponds to the significance level that was specified by the analyst performing the test. A size that is too large means that false detection of a climate change signal that is not there will occur more often than indicated by the specified significance level, while one that is too small means an increased risk of failing to detect a signal that is there. A test with high power has a high chance of being able to detect climate change signals from internal climate variability, if they exist. The length of data records, the strength of the signal relative to internal variability, and the significance level affect the power of a test. The first two factors are fixed in applications, and thus we consider whether the power of field detection can be optimized by adjusting the significance levels of the local and field tests.

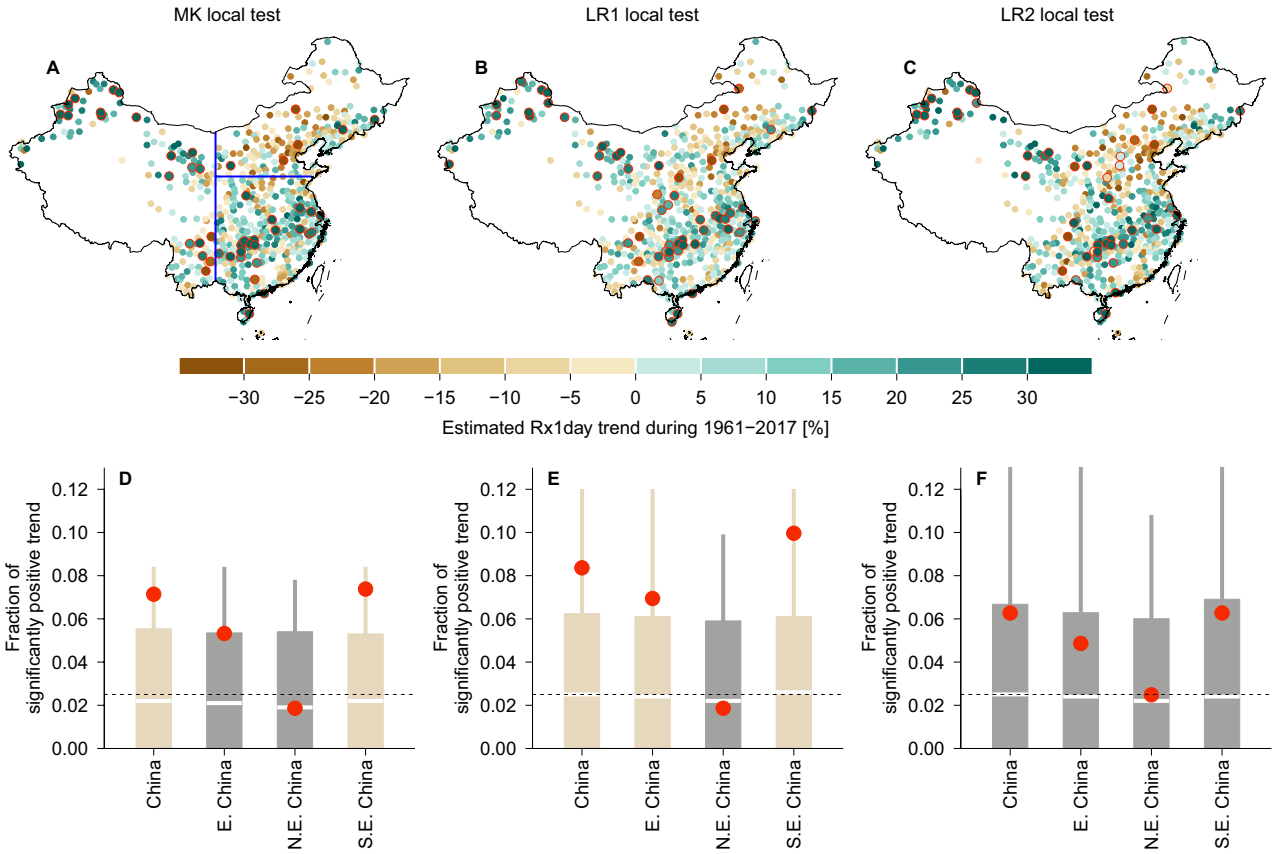

Figure 1. Field significant tests built with different local tests implemented on observations of Rx1day. (a–c) Tends in Rx1day during 1961–2017 estimated by Sen's slope (a), a GEV distribution with location parameter that varies linearly with time (b), and a GEV distribution in which the log of the scale parameter varies linearly with time (c). Stations with significant trends at a local significance level of 2.5% are marked by red circles. Trends are expressed as percentage changes over the 1961–2017 period relative to the medians of that period. (d and e) Fractions of significantly positive trends in the Rx1day observations of different regions (red dots) determined by MK (d), LR1 (e), and LR2 (f) local tests at a specified local significance level of 2.5% and the corresponding fractions under internal variability as inferred by space-time block bootstrapping (boxplots). White lines, bars, and whiskers show the 50th, 95th, and 99th percentiles of the fractions under internal variability, respectively. Dashed black lines mark the specified local significance level of 2.5%. Regions where the intensification of Rx1day is field significant at a field significance level of 5% are marked in light yellow and otherwise in light gray. Geographic definitions of the studied regions are shown in (a), that is, the whole China, eastern China, northeast China, and southeast China.

4. Results and Discussion

4.1. The Dependence of Field Detection on Local Detection Methods

Figure 1 presents the results of field significance tests with MK, LR1, and LR2 local tests at a specified local significance level of 2.5% for positive trends on the 1961–2017 Rx1day observations over China. We see an overall pattern of increases in southeast and northwest China and decreases in northeast China regardless of the local detection method used (Figures 1a–1c), consistent with the "wetting in the south and drying in the north" pattern that has been found in changes in both mean and extreme precipitation in China (e.g., Dong et al., 2020; Li et al., 2017, 2018). Despite that, there are obvious differences in the number of stations showing significant trends and their magnitudes evaluated by different local detection methods.

As a result, inconsistent field significance conclusions can be drawn with different local detection methods (Figures 1d–1f). For example, the fraction of significantly positive trends in the Rx1day observations over the country falls well above the 95th percentile of the corresponding fractions expected when forced trends are absent by construction based on the MK and LR1 tests (Figures 1d and 1e), while this is not the case according to the LR2 test (Figure 1f). It thus follows that the intensification of Rx1day at the national scale can be claimed to be either field significant at a field significance level of 5%, or not, depending on the choice of local detection method. Similar findings are applicable to other regions except northeast China

where there is an overall drying tendency. Conducting one-sided tests for negative trends similarly produces inconsistent conclusions in this region (not shown). Overall, these results highlight that conclusions about whether extreme precipitation is intensifying in China are sensitive to the choice of local detection method. The discrepancies are important from a policy perspective. Policy makers are trying to understand whether there is robust evidence of changes in precipitation regimes within their jurisdictions that require a policy or adaptation response. It is therefore important to understand why seemingly equally defensible testing approaches produce conflicting answers.

Information about the null distribution of the field test statistic (boxplots in Figures 1e and 1f), that is, the (false) rejection rate across stations in a stationary climate, informs which local detection method is preferred. When the data are serially independent, this false rejection rate should be statistically indistinguishable from the specified local significance level, such as 2.5% in the present analysis. This is roughly the case for all tests in all regions that we consider (dashed black lines vs. white solid lines in Figures 1d–1f). But, the distribution of the false rejection rates is wider for the two LR tests than for the MK test in all regions (boxplots in Figures 1d–1f), suggesting that the use of the LR tests results in higher uncertainty in the determination of field significance. This can be induced by several factors such as (1) whether the annual block size from which Rx1day is extracted is long enough to ensure that the distribution of the block maxima is reasonably well approximated by a GEV distribution, (2) whether the data sample size is large enough to ensure the convergence of the likelihood ratio statistic to a chi-square distribution (e.g., Coles, 2001), (3) whether GEV model being used in the LR test is overly complex, and (4) and whether the likelihood maximization procedure for GEV estimation converges to the correct solution when trends are not present. In contrast, the MK test requires simpler assumptions and no iterative likelihood estimation of parameters, resulting in a narrower null distribution with which to assess field significance.

4.2. The Size and Power of Field Detection

To gain further insight into the choice between MK and LR tests, Figure 2 contrasts the size and power of field significance tests with MK and LR1 local tests for positive trends in the constructed simulations. We find that the size of local test is close to the specified local significance level of 2.5% over the majority of grid cells for both MK and LR1 tests (Figures 2a and 2b). The size of field test also does not seem to depend on the choice between MK and LR1 tests (Figure 2c). For field test at the national scale, the size is generally consistent with the specified field significance level of 5% for both tests, but it tends to be markedly smaller than the specified level at subnational scales (Figure 2c). This may be because the field test statistic, which is the fraction of local tests at which rejection of the null hypothesis occurs, can only take a limited number of different values in smaller regions. In an extreme case, for example, a region with only 10 grid cells and Rx1day observations in these grid cells are spatially independent, only 11 local rejections fractions are possible (i.e., 0, 0.1, 0.2, ..., 1). Thus, it may not be possible to identify a critical value for the field test statistic that corresponds closely to a desired field significance level, such as 5% (see Figure S2 for a detailed illustration). In real-world applications, Rx1day data are not independent in space, and thus the number of effectively independent local tests can be substantially smaller than the number of stations or grid cells in a region. This highlights the need for there to be sufficient "*independent*" local tests for reliably detecting changes to climate extremes in small regions and/or with strong spatial dependence.

We find that the MK test exhibits comparable or higher power of local detection than LR1 test (Figures 2d and 2e), and that the power of field detection with the MK test is consistently higher than with the LR test (Figure 2f). We also see that the power of field detection declines with the size of region, partly because the role of internal variability is large in small regions and/or the effective number of independent local tests is small. The anticipated form of the trend when the null hypothesis is false affects the power of a test by affecting how evidence counter to the null hypothesis is assessed. In the LR1 test, the alternative hypothesis to the null hypothesis is specified by a GEV distribution with a location parameter that is linearly dependent on time, and thus the LR statistic compares the fit of stationary GEV models to models with this specific form of trend. If the actual trend has a different form, it will likely be more difficult to detect. A more complex model for the trend, such as a GEV distribution that also allows its scale parameter to vary with time as in the LR2 test, may be able to better accommodate the actual trend, but may not lead to higher power, because more parameters must be estimated from the same finite data resource, and because convergence

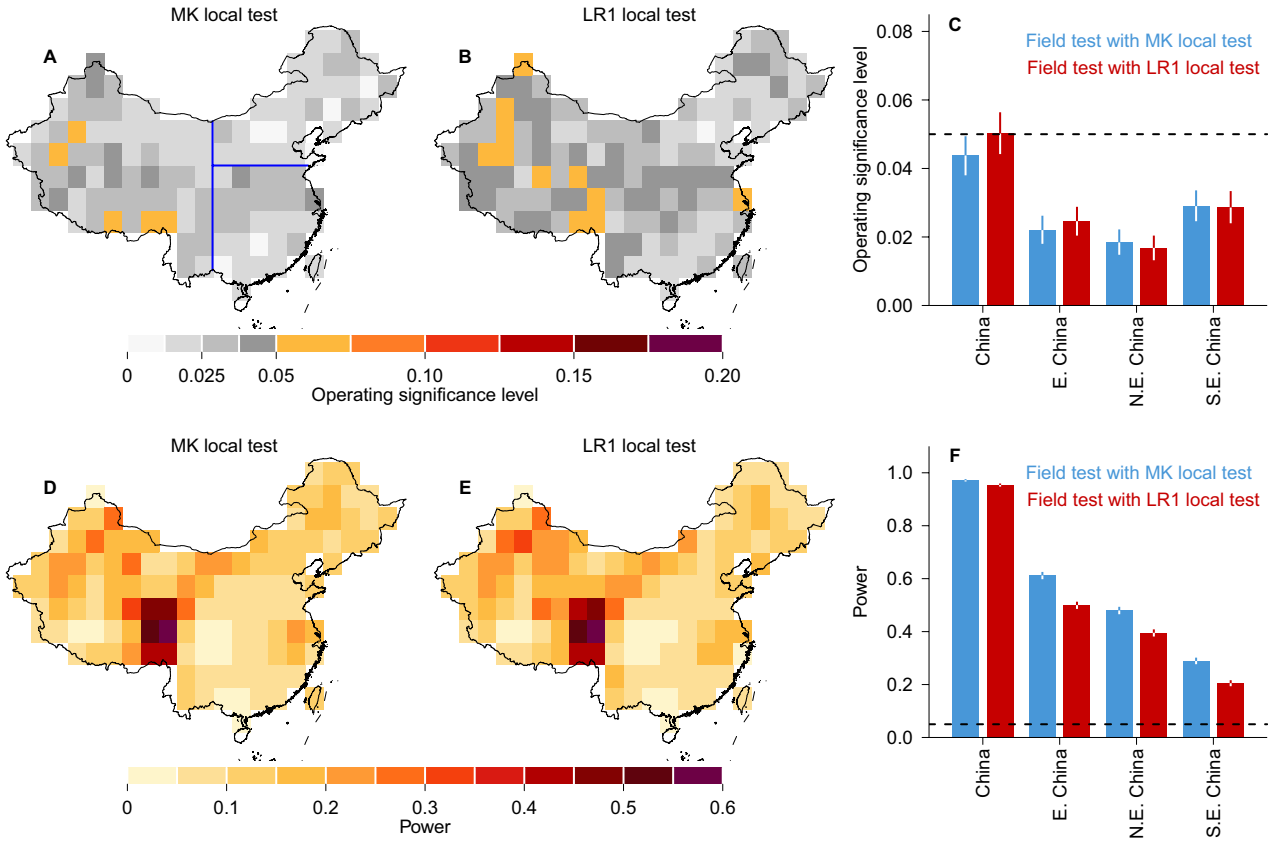

Figure 2. The size and power of different field significance tests estimated with reconstructed simulations of Rx1day. (a and b) Estimates of the size of MK (a) and LR1 (b) local tests at a local significance level of 2.5% for individual grid cells. (c) Estimates of the size of field tests using MK (blue) and LR1 (red) local tests at a local significance level of 2.5% and a field significance level of 5% for different regions. (d–f) As in (a–c) but for estimates of the power the local and field tests. In (c and f), bars show the best estimates, while whiskers show central 95% uncertainty ranges of the estimates. Dashed black lines mark the specified field significance level of 5%. Geographic definitions of the regions are shown in (a).

of the LR statistic to its asymptotic chi-square distribution will occur more slowly. This is confirmed by the further reduced power of field detection when the LR2 test was used (not shown). In contrast, the simple MK test uses an alternative hypothesis that trend, if present, is exhibited by a monotone tendency upwards or downwards without specifying a particular form for that tendency. Thus, it is not surprising to see the more powerful field detection with the MK local test.

4.3. Optimizing the Power of Field Detection

We now consider whether the power of the MK-based field significance test can be optimized by adjusting the local and field significance levels. We first conducted a set of field tests for the intensification of Rx1day in station observations for different periods and different regions using a fixed field significance level of 5% but varying local significance levels. Figure 3 presents the ratios of the field significance level to the *p* values of different field tests. Field significance at the 5% level can be claimed if this ratio is greater than 1. We find that the performance of the field test is not all that sensitive to the choice of local significance level, implying that increasing the power of local detection by using a larger local significance level will not necessarily increase the power of field detection. We confirmed that for all the tests, the operating local significance levels are consistent with the specified values on average over the country and are relatively uniformly distributed in space (e.g., Figure S3), suggesting that they are reliable and the field tests are not biased as they would be if the operating local significance levels are spatially uneven.

As expected, increasing the significance level of field test can increase the power of field detection (Figure S4), but at the cost of increasing the chance of false field significance detections when there is actually

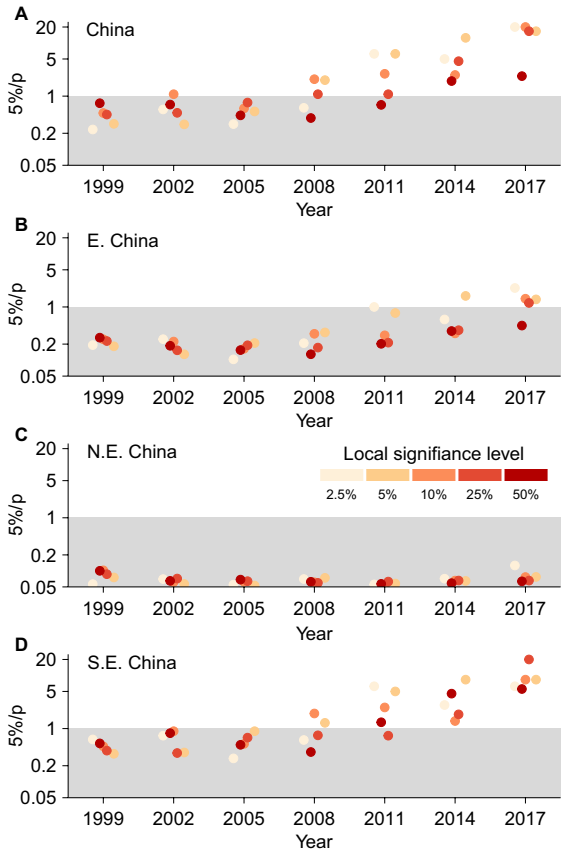

Figure 3. Field significance tests of observed changes in precipitation extremes with varying local significance levels. Panels show ratios of the specified field significance level (i.e., 5%) to the *p* values of different field significance tests with MK local tests at different local significance levels for the intensification of Rx1day in China (a), eastern China (b), northeastern China (c), and southeastern China (d) for different periods from 1961. The last year of the period considered is marked along the horizontal axes. Geographic definitions of the regions can be found in Figure 1a.

no climate change. Nevertheless, given the large body of observational and modeling evidence for the intensification of extreme precipitation over most landmasses with global warming (e.g., Allen & Soden, 2008; Zhang et al., 2013), earlier detection enabled by a more powerful test is desired since it allows an earlier societal response that may help avoid increased risks.

4.4. Emergence of Intensification Signals of Extreme Precipitation in China

A recent study reported that extreme precipitation changes in China had not emerged from internal climate variability by the year 2012 in observations, and would be unlikely to emerge before around 2035 based on climate model simulations driven by the representative concentration pathway 8.5 forcing scenario (Li et al., 2018). This study employed a LR local test that involves an extended GEV distribution for the 3 largest daily precipitation accumulations in a year, comparing distributions with time-varying location and scale parameters to distributions with constant parameters. Our study, however, implies that such a field test may be less powerful than a field test using the simple MK trend diagnostic.

Using a field test based on MK tests with a local significance level of 10%, we find that the overall intensification of Rx1day in China from 1961 became field significant at the 5% field significance level in 2008 (Figure 3a). The intensification became field significant by 2014 regardless of the choice of local significance level. Similar conclusions can be drawn for southeastern China (Figure 3d). As Rx1day in northeastern China is decreasing (Figure 1) due to the weakened northward moisture transport and convergence by decreasing Asian summer monsoon over the past decades (e.g., Ding et al., 2008, 2009), no signal of extreme precipitation intensification is detected there (Figure 3c). Nevertheless, field significant intensification of Rx1day is detected over eastern China using the full record ending in 2017 (Figure 3b).

5. Conclusions

To summarize, our study clarifies that the widely used field significance detection method for changes in annual extremes of climate variables, which is built with a local likelihood ratio test for a generalized extreme value distribution with time-varying parameters against one with constant parameters, is generally not as powerful as simply using the Mann-Kendall test for an overall tendency. This is mainly because it involves several assumptions that may not be fully satisfied in applications to real extreme precipitation records. The study also clarifies that although increasing the significance level of the local tests can increase the power of local detection, doing so will not strongly affect the power of field detection. The power of field detection can, however, be increased by increasing the significance level of field test, but at the cost of also increasing the chance of false detection of field significance in the absence of a climate change signal. These lessons allow us to report, with enhanced confidence, that the intensification of Rx1day that has occurred in China since 1961 became field significant early this century, which is substantially earlier than previously reported. Our results provide useful insights into why seemingly equally defensible field significance detection approaches to climate change produce conflicting answers.

Data Availability Statement

The observations used in this study are downloaded from https://data.cma.cn/data, while the simulations can be available from https://esgf-node.llnl.gov/search/cmip6. R codes for implementing the field test methods are available as online supporting materials.

Acknowledgments
We thank the Canadian Center for Climate Modeling and Analysis of Environment and Climate Change Canada for executing and making available the CanESM5 large ensemble simulations. This study was supported by the National Key R&D Programs of China (2018YFC1507700). Jianyu Wang and Chao Li was also supported by the National Natural Science Foundation of China (42075026).

References

Alexander, L. V., Zhang, X., Peterson, T. C., Caesar, J., Gleason, B., Klein Tank, A. M. G., et al. (2006). Global observed changes in daily climate extremes of temperature and precipitation. *Journal of Geophysical Research*, *111*, D05109. https://doi.org/10.1029/2005JD006290

Allen, R. P., & Soden, B. J. (2008). Atmospheric warming and the amplification of precipitation extremes. *Science*, *321*, 1481–1484. https://doi.org/10.1126/science.1160787

Chen, Y., Li, W., Jiang, X., Zhai, P., & Luo, Y. (2021). Detectable intensification of hourly and daily scale precipitation extremes across eastern China. *Journal of Climate*, *34*, 1185–1201. https://doi.org/10.1175/JCLI-D-20-0462.1

Coles, S. G. (2001). *An introduction to statistical modeling of extreme values*. Springer.

Ding, Y., Sun, Y., Wang, Z., Zhu, Y., & Song, Y. (2009). Inter-decadal variation of the summer precipitation in China and its association with decreasing Asian summer monsoon Part II: Possible causes. *International Journal of Climatology*, *29*, 1926–1944. https://doi.org/10.1002/joc.1759

Ding, Y., Wang, Z., & Sun, Y. (2008). Inter-decadal variation of the summer precipitation in East China and its association with decreasing Asian summer monsoon. Part I: Observed evidences. *International Journal of Climatology*, *28*, 1139–1161. https://doi.org/10.1002/joc.1615

Dong, S., Sun, Y., & Li, C. (2020). Detection of human influence on precipitation extremes in Asia. *Journal of Climate*, *33*, 5293–5304. https://doi.org/10.1175/JCLI-D-19-0371.1

Fischer, E. M., & Knutti, R. (2014). Detection of spatially aggregated changes in temperature and precipitation extremes. *Geophysical Research Letters*, *41*, 547–554. https://doi.org/10.1002/2013GL058499

Fischer, E. M., Sedláček, J., Hawkins, E., & Knutti, R. (2014). Models agree on forced response pattern of precipitation and temperature extremes. *Geophysical Research Letters*, *41*, 8554–8562. https://doi.org/10.1002/2014GL062018

Fischer, R. A., & Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest members of a sample. *Proceeding of the Cambridge Philosophical Society*, *24*, 180–190. https://doi.org/10.1017/S0305004100015681

Kendall, M. G. (1975). *Rank correlation methods*.London: Griffin.

Kiktev, D., Sexton, D. M. H., Alexander, L., & Folland, C. K. (2003). Comparison of modeled and observed trends in indices of daily climate extremes. *Journal of Climate*, *16*, 3560–3571. https://doi.org/10.1175/1520-0442(2003)016<3560:COMAOT>2.0.CO;2

Leadbetter, M. R, Lindgren, G, & Rootzen, H (1983). *Extremes and related properties of random sequences and processes*. New York Springer-Verlag Inc. https://doi.org/10.1007/978-1-4612-5449-2

Li, C., Zwiers, F., Zhang, X., & Li, G. (2019). How much information is required to well constrain local estimates of future precipitation extremes? *Earth's Future*, *7*, 11–24. https://doi.org/10.1029/2018EF001001

Li, C., Zwiers, F., Zhang, X., Li, G., Sun, Y., & Wehner, M. (2021). Changes in annual extremes of daily temperature and precipitation in CMIP6 models. *Journal of Climate*, *34*, 3441–3460. https://doi.org/10.1175/JCLI-D-19-1013.1

Li, H., Chen, H., & Wang, H. (2017). Effects of anthropogenic activity emerging as intensified extreme precipitation over China. *Journal of Geophysical Research: Atmospheres*, *122*, 6899–6914. https://doi.org/10.1002/2016JD026251

Li, W., & Chen, Y. (2020). Detectability of the trend in precipitation characteristics over China from 1961 to 2017. *International Journal of Climatology, 41*(Suppl. 1), E1980–E1991. https://doi.org/10.1002/joc.6862

Li, W., Jiang, Z., Zhang, X., & Li, L. (2018). On the emergence of anthropogenic signal in extreme precipitation change over China. *Geophysical Research Letters*, *45*, 9179–9185. https://doi.org/10.1029/2018GL079133

Li, Z., Yan, Z., Tu, K., & Wu, H. (2015). Changes of precipitation and extremes and the possible effect of urbanization in the Beijing metropolitan region during 1960-2012 based on homogenized observations. *Advances in Atmospheric Sciences*, *32*, 1173–1185. https://doi.org/10.1007/s00376-015-4257-x

Livezey, R. E., & Chen, W. Y. (1983). Statistical field significance and its determination by Monte Carlo techniques. *Monthly Weather Review*, *111*, 46–59. https://doi.org/10.1175/1520-0493(1983)111<0046:SFSAID>2.0.CO;2

Lorenz, R., Stalhandske, Z., & Fischer, E. M. (2019). Detection of a climate change signal in extreme heat, heat stress, and cold in Europe from observations. *Geophysical Research Letters*, *46*, 8363–8374. https://doi.org/10.1029/2019GL082062

Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica*, *13*, 245–259. https://doi.org/10.2307/1907187

Risser, M. D., Paciorek, C. J., O'Brien, T. A., Wehner, M. F., & Collins, W. D. (2019). Detected changes in precipitation extremes at their native scales derived from in situ measurements. *Journal of Climate*, *32*, 8087–8109. https://doi.org/10.1175/JCLID-D-19-0077

Storch, H. v., & Zwiers, F. W. (1984). *Statistical analysis in climate research*. Cambridge University Press. https://doi.org/10.1017/CBO9780511612336

Sun, Q., Zhang, X., Zwiers, F., Westra, S., & Alexander, L. V. (2021). A global, continental, and regional analysis of changes in extreme precipitation. *Journal of Climate*, *34*, 243–258. https://doi.org/10.1175/JCLI-D-19-0892.1

Swart, N. C., Cole, J. N. S., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P., et al. (2019). The Canadian Earth System Model version 5 (CanESM5.0.3). *Geoscientific Model Development*, *12*, 4823–4873. https://doi.org/10.5194/gmd-12-4823-2019

Wang, X. L., & Swail, V. R. (2001). Changes of extreme wave heights in northern hemisphere oceans and related atmospheric circulation regimes. *Journal of Climate*, *14*, 2204–2221. https://doi.org/10.1175/1520-0442(2001)014<2204:COEWHI>2.0.CO;2

Westra, S., Alexander, L. V., & Zwiers, F. W. (2013). Global increasing trends in annual maximum daily precipitation. *Journal of Climate*, *26*, 3904–3918. https://doi.org/10.1175/JCLI-D-12-00502.1

Wilks, D. S. (2006). On "field significance" and the false discovery rate. *Journal of Applied Meteorology and Climatology*, *45*, 1181–1189. https://doi.org/10.1175/JAM2404.1

Zhang, W., & Zhou, T. (2019). Significant increases in extreme precipitation and the associations with global warming over the global land monsoon regions. *Journal of Climate*, *32*, 8465–8488. https://doi.org/10.1175/JCLI-D-18-0662.1

Zhang, X., Wan, H., Zwiers, F. W., Hegerl, G. C., & Min, S. K. (2013). Attributing intensification of precipitation extremes to human influence. *Geophysical Research Letters*, *40*, 5252–5257. https://doi.org/10.1002/grl.51010

Zhang, X., Wang, J., Zwiers, F. W., & Groisman, P. Y. (2010). The influence of large-scale climate variability on winter maximum daily precipitation over North America. *Journal of Climate*, *23*, 2902–2915. https://doi.org/10.1175/2010JCLI3249.1

Zhang, X., Wang, K., & Boehrer, B. (2020). Variability in observed snow depth over China from 1960 to 2014. *International Journal of Climatology*, *41*, 374–392. https://doi.org/10.1002/joc.6625

Zhang, X., & Zwiers, F. W. (2004). Comment on "Applicability of prewhitening to eliminate the influence of serial correlation on the Mann-Kendall test" by Sheng Yue and Chun Yuan Wang. *Water Resources Research*, *40*, W03805. https://doi.org/10.1029/2003WR002073

References From the Supporting Information

Yue, S., & Wang, C. Y. (2002). Applicability of prewhitening to eliminate the influence of serial correlation on the Mann-Kendall test. *Water Resources Research*, *38*(6), 1068. https://doi.org/10.1029/2001WR000861

Zwiers, F. W., & von Storch, H. (1995). Taking serial correlation into account in tests of the mean. *Journal of Climate*, *8*, 336–351. https://doi.org/10.1175/1520-0442(1995)008<0336:TSCIAI>2.0.CO;2