Identifying Significant Trends in Heat Wave Humidity with Multiple Comparison Corrections

Tristan Ballard
Stats 205: Introduction to Nonparametric Statistics

1. Introduction

Improving our understanding of how heat waves will respond to climate change is critical for adequate planning and adaptation. The impacts of heat waves can be widespread across a variety of sectors, making it a particularly important extreme event to study in the context of climate change. In 2003, a heat wave in Europe killed over 70,000 people. The prolonged Russian heat wave in 2010 killed roughly 54,000 (Perkins, 2015). Since crops can be susceptible to extreme temperatures as well, heat waves pose a threat to the agricultural industry and global food security. Moreover, rapid increases in demand for water resources and energy during these events puts pressure on already strained infrastructure, especially in urban centers (Habeeb et al., 2015). With evidence that heat waves have already begun to intensify (Frich et al., 2002; Della-Marta et al., 2007; Perkins et al., 2012), it is essential that we continue our research into heat wave dynamics. Here I focus on the contributions of humidity to extreme heat wave events.

While temperature is the key determinant of heat wave severity, humidity has been shown to play a key role in their intensity and physiological effects, with direct links to human health and safety (Sherwood and Huber, 2010). High humidity inhibits the body's ability to sweat, increasing the risk of adverse health effects (Gasparrini and Armstrong, 2011). Pal and Eltahir (2015) used climate models to simulate wet-bulb temperature, a metric for temperature that incorporates humidity. They found that certain regions of the Middle East may become uninhabitable by the end of the century due to heat effects intensified by high humidity.

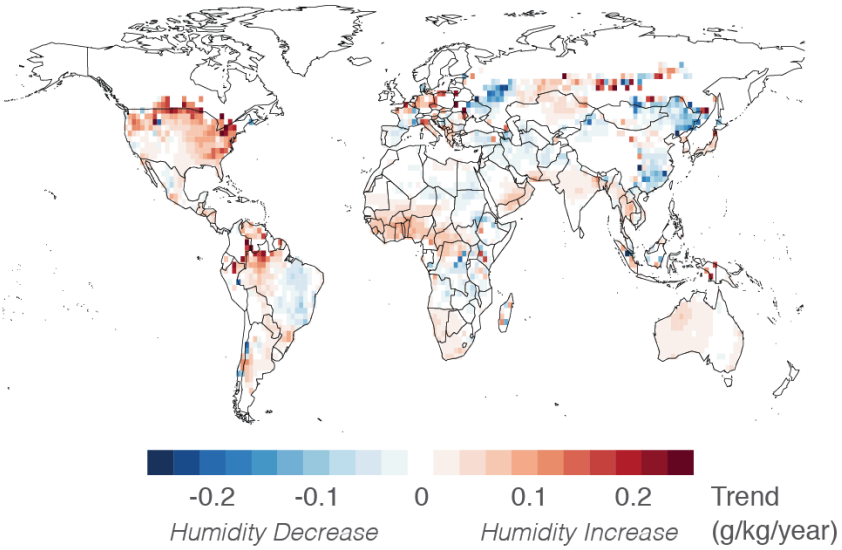

Figure 1. Trend in specific humidity during heat wave events, defined as periods where maximum temperature exceeds 35°C (a metric different than the one used in this analysis).

Despite the importance of humidity to human health, there is still significant uncertainty over how it will respond regionally and globally to climate change, and consequently how its role in heat waves may change (Willett et al., 2014). **I hypothesize that the probability of extreme heat wave events, characterized by both high temperatures and humidity, has increased over particular regions of the globe.** There are already signs that humidity during heat wave events has increased, in particular over the eastern U.S. (Fig. 1).

Here as an initial exploration of this hypothesis I calculate the trend in humidity during historical heat waves, as in Figure 1. While the significance of these trends can be calculated individually with their respective p-values, this does not take into account the issue of testing many hypotheses. I therefore adjust the p-values of calculated trends and reevaluate their significance using several correction methods described in Section 3.

2. Data

Historical data are from the NCEP-DOE Reanalysis II dataset (Kanamitsu et al., 2002). This dataset provides maximum temperature and specific humidity estimates with complete spatial coverage on a global 192x94 grid (Fig. 2). This dataset and other 'reanalysis' datasets are commonly used in climate research for historical data. However, it is not technically observational data. The dataset incorporates observations of numerous variables from weather stations, satellites, and other sources into a weather model that interpolates the missing observations both spatially and temporally. In general the data has higher fidelity over the U.S., Europe, and Australia than in other parts of the world.

Each map of temperature or humidity estimates are available at 6hr intervals from 1979-2014, leaving a total of 36 years * 365 days * 4 maps/day = 52,560 maps for each variable. Figure 2 shows the average humidity (g/kg) during June, computed as the average of the 6hr interval maps for June across all years. This gives a visual idea of the spatial resolution.

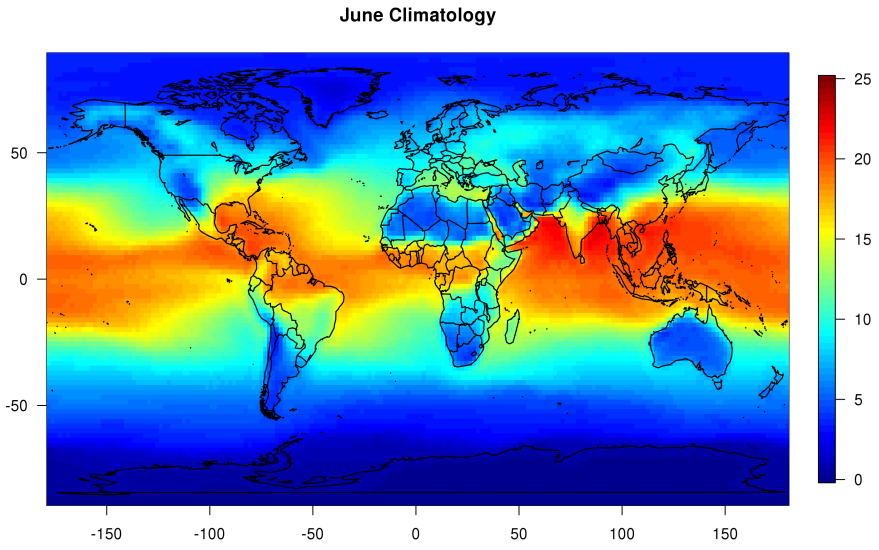

Figure 2. Average June humidity. The warmest areas of the globe also tend to be the most humid, but there is considerable daily and spatial variability (not shown). How humidity has changed in the past can help inform predictions for the future.

I define heat waves following the convention proposed in Perkins and Alexander (2013), where a heat wave is any period with 3 or more consecutive days with the maximum temperature exceeding the 90th percentile for that date and grid cell. The distribution from which to calculate the percentiles is compiled from the observed temperature values on that calendar date for that particular spatial grid cell as well as the observed values at that grid cell 7 days following and 7 days preceding that date. Note that this definition will indicate 'warm spells' and not necessarily heat waves, since you will have values for days in winter as well. I present here the calculations for January and July only, representing a summer month for the Southern Hemisphere and Northern Hemisphere, respectively.

3. Adjusting p-values with multiple comparisons corrections

3.1 Motivation

When using p-values to determine the significance of a test statistic, one rejects the null hypothesis when the p-value is below a predetermined threshold, usually somewhere between .01 and .10. However, there exists the possibility that one incorrectly rejects the null. If running 1000 tests, for example, one would expect on average 50 'significant' test results rejecting the null even if the null were indeed true. These are known as false positives, or Type I error. The data here requires around 6,000 tests (ocean values are masked). Numerous methods have been developed to try to reduce this Type I error rate, controlling either the family-wise error rate (FWER) or the false discovery rate (FDR).

The FWER is the probability of having at least one false positive test result. This probability is $1 - P(no\ rejections) = 1 - (1 - \alpha)^n$ where n is the number of tests and α is the p-value threshold. With large n the FWER is almost inevitably 1, so it is not logical to control this metric. Plus with 6,000 tests it is acceptable scientifically to have 1 or more false positives; one just wants the relative number of false positives to be small. This leads to the FDR, which is the expected proportion of false positives relative to the total number of tests rejected (the false discovery proportion). In statistics the expectation of a proportion is often referred to as a rate. The primary focus in this paper is on methods that seek to limit the FDR.

3.2 Benjamini and Hochberg (BH) correction

The Benjamini and Hochberg (BH) correction is one of the most widely used methods for controlling the FDR and is relatively straightforward (Benjamini and Hochberg, 1995). Assuming n hypothesis tests were computed, first order the observed n p-values $p_1 \ldots p_n$ in increasing order. Then, find the largest i such that $p_i \leq \frac{i}{n}\alpha$ for a given α. Reject the null (i.e. mark as statistically significant) the tests corresponding to $p_1 \ldots p_i$. Under this method, which assumes the tests are independent, the FDP will on average be $\leq \alpha$. Selecting smaller values of α will yield fewer rejections of the null and thus more conservative results. Throughout the analysis I set $\alpha = .05$. Note as well that as the number of tests considered, n, increases, the threshold p-value for significance, p_i, becomes smaller.

3.3 Benjamini and Yekutieli (BY) correction

The Benjamini and Yekutieli (BY) correction is a modification to the BH method that is more conservative (Benjamini and Yekutieli, 2001). The original BH method is suitable when the test statistics are independent; however, in practice test statistics are rarely completely independent. In this study, for example, there is likely some spatial correlation. The authors do show, however, that the BH method is appropriate if the test statistics are normally distributed and positively correlated. The BY method was developed for other possible settings of dependence between test statistics, for example negative correlation, and thus has far less strict assumptions.

The setup for BY is the same as for BH with the exception that one finds the largest *i* such that $p_i \leq \frac{i}{nm}\alpha$ where $m = \sum_{j=1}^{n} \frac{1}{j}$. Since *m* is always greater than 1 for *n* greater than 1, the threshold p-value for significance, p_i, becomes smaller than in BH and therefore yields fewer statistically significant hypothesis tests.

3.4 Ventura correction

The Ventura correction is a modification to the BH correction that the authors argue is better suited for data with spatial correlation. The setup is the same as for BH except that one finds the largest i such that $p_i \leq \frac{i}{n}\alpha(1-m)$ where $m = n_{H_A}/n$, the ratio of the (unknown) number of true alternative hypotheses to the total number of hypotheses. The justification is that with the BH procedure, the actual upper bound on the FDR is $(1-m)\alpha$. Since that value will be $\leq \alpha$ this provides a tighter upper bound than α. In practice m needs to be estimated from the data, and the authors discuss different ways of estimating this empirically and their final recommendation. Overall, similar to the BY correction, since $0 \leq m \leq 1$, the result will be a more conservative correction method than BH.

The authors demonstrate their method and compare it with BH and BY using simulated data with spatial correlation. They conclude that BY is too conservative a method, so BH should be used in general on spatial data and that their modified BH performs best. Wilks (2006) reviewed the multiple comparisons problem in the context of spatial data and recommended the BH and Ventura corrections, although his support for the Ventura correction in place of BH was tepid. He suggests that it may be possible to improve on BH in the way Ventura suggests but that with his simulated data the Ventura algorithm for estimating m was inaccurate and BH outperformed.

3.5 Bonferroni correction

Unlike the other correction methods presented so far, the Bonferroni correction aims to control the FWER instead of the FDR. Instead of setting a significance threshold for p-values of α, the threshold is modified by simply dividing by the total number of tests. The new p-value threshold is therefore $\frac{\alpha}{n}$.

While the Bonferroni correction is the simplest of the methods for controlling FWER, it is extremely conservative especially for large n and is now relatively outdated. Newer methods for controlling the FWER based in part on the Bonferroni correction have been developed that are less conservative, such as the Holm, Hochberg, or Hommel corrections (Holm, 1979; Hochberg, 1988; Hommel, 1988). However, control of the FWER is primarily appropriate when testing only a handful of hypotheses; none of these are appropriate in this analysis containing thousands of tests. I include the Bonferroni correction for comparison and demonstration purposes only. Though not shown, I ran the analysis with the Holm correction as well and it was only slightly less conservative than the original Bonferroni correction.

3.6 Debate over multiple comparison correction methods

Deciding which correction method is best suited for one's analysis can be difficult because each test has its own assumptions and degree of strictness. Sometimes expert judgment and intuition can be used to decide how conservative a correction method is reasonable. An overly strict method will reduce the false positive (Type I) error rate at the expense of increasing the false negative (Type II) error rate. However, the relative cost of each error rate will be problem specific. For example, in medical studies the high cost of false positives (e.g. falsely claiming a drug cures cancer) may justify using very conservative methods.

Interestingly, apart from debate over which method to use is a larger debate over whether these correction methods should be used at all. Rothman (1990) makes the case against using multiple comparison corrections. Primarily, he criticizes the idea of a universal null hypothesis, the assumption from the beginning that all the hypotheses are true and we must have significant evidence to reject from particularly extreme observed test statistics. He claims this is an unreasonable assumption because in reality nature follows known laws; observations are rarely truly random.

More recently, Gelman et al. (2012) argue similar points as Rothman about the faults in assuming a universal null hypothesis. They instead (not surprisingly) develop a Bayesian method as a panacea to the debate. After reading these articles and some discussions online I think the criticisms of the universal null hypothesis really just parallel criticisms of using p-values in the first place. Those concerns are valid but do not negate the utility of p-values. P-values are still tremendously prevalent in research, so if we are to use them in making decisions we should use the corresponding correction methods as well. Though not a scientific measure of where experts stand on this issue, the BH paper has over 32,000 cites since 1995 while the Rothman paper has only 2,500 since 1990.

4. Results

I calculate trends in humidity during heat wave events for January and July globally and apply each of the correction methods mentioned above. The Ventura correction is applied with two thresholds q=0.05 and q=0.01. Note that for Ventura I use q in place of α. Presented first are global results then the results for just the U.S. and Canada.

4.1 Global humidity trends

Global results are shown in Figures 3 and 4. In interest of brevity I will focus discussion on July. As seen in the unmodified humidity trend plot (Fig. 3, column 1), trends are variable across the Northern Hemisphere with a notable positive trend over the U.S. and Canada. This region, shown by the mass of red pixels, remains after applying all the correction methods, suggesting that this is not a case of false positives.

Instead, the majority of identified false positives were for trend values whose magnitudes were originally quite small. Those areas are shown in purple in the maps below. Applying each correction method changes some of these purple pixels to white, meaning the correction method adjusted their p-values to be non-significant.

There is variability in how strict/conservative each method was with this set of hypothesis tests (Table 1). As expected, the BY correction was more conservative than the BH correction, shown by fewer p-values deemed significant in the BY map after applying the adjustment. Similarly, the Ventura correction with q=.01 was far more conservative than with q=.05. The Bonferroni correction was the most conservative by far. Interestingly, the Ventura correction with q=.05 had practically no effect on the results and for July it did not change the results at all from the original p-value analysis.

One obvious pattern is that the majority of 'false positives' were in Antarctica, not the regions where we would expect heat waves in the first place. Therefore, I reran the analysis focusing solely on the U.S. and Canada in July, with results presented in the next section.

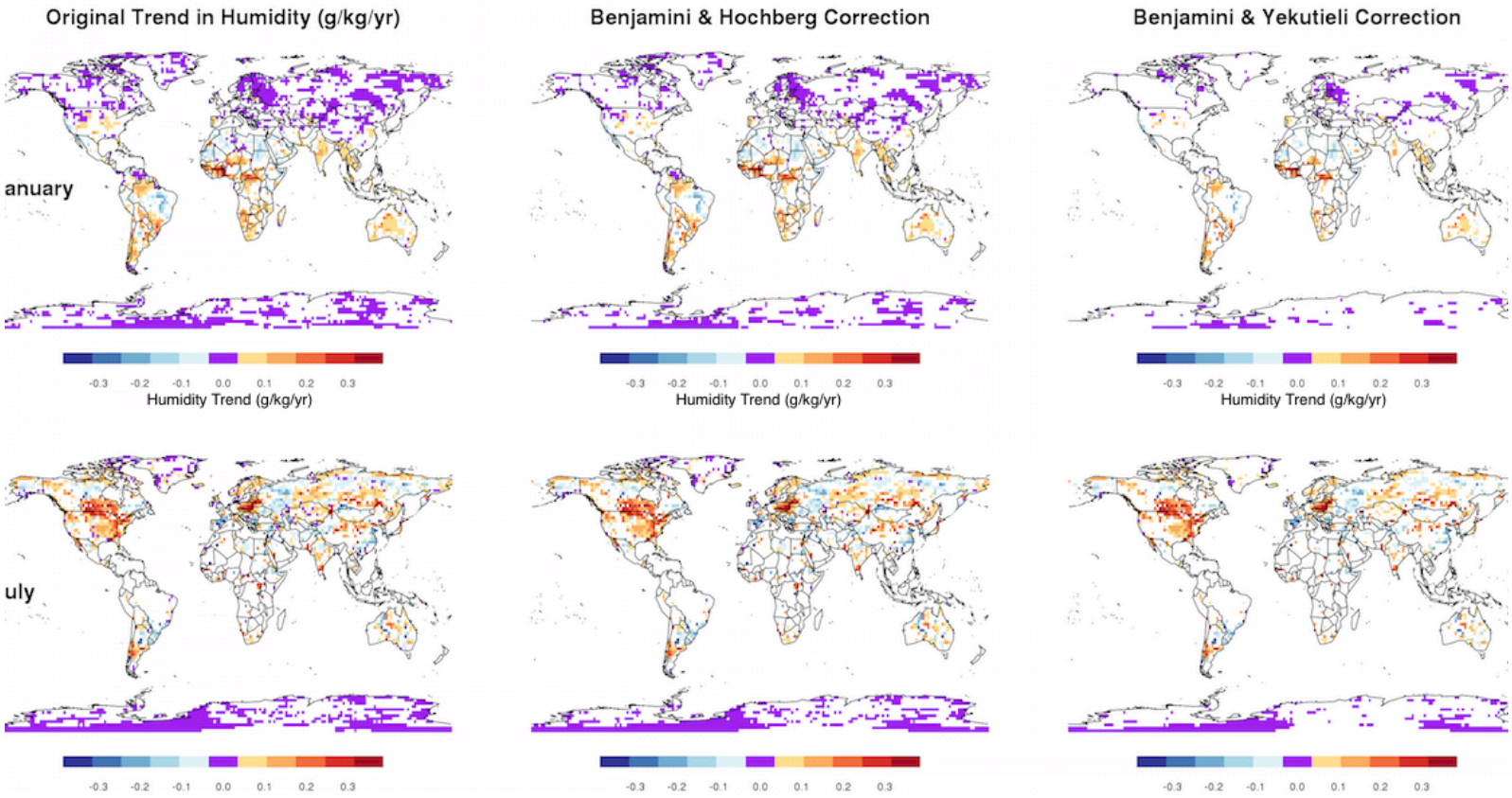

Figure 3. Trend in humidity during heat waves from 1979-2014 (Column 1). Two correction procedures are applied in Columns 2 and 3. Only land areas were analyzed. White over land corresponds to areas where the humidity trend was not significant. Purple regions are still significant but have the smallest trend values, highlighted in purple to aid visually that the correction methods primarily affect these smaller trends and not the regions of dark red or blue.

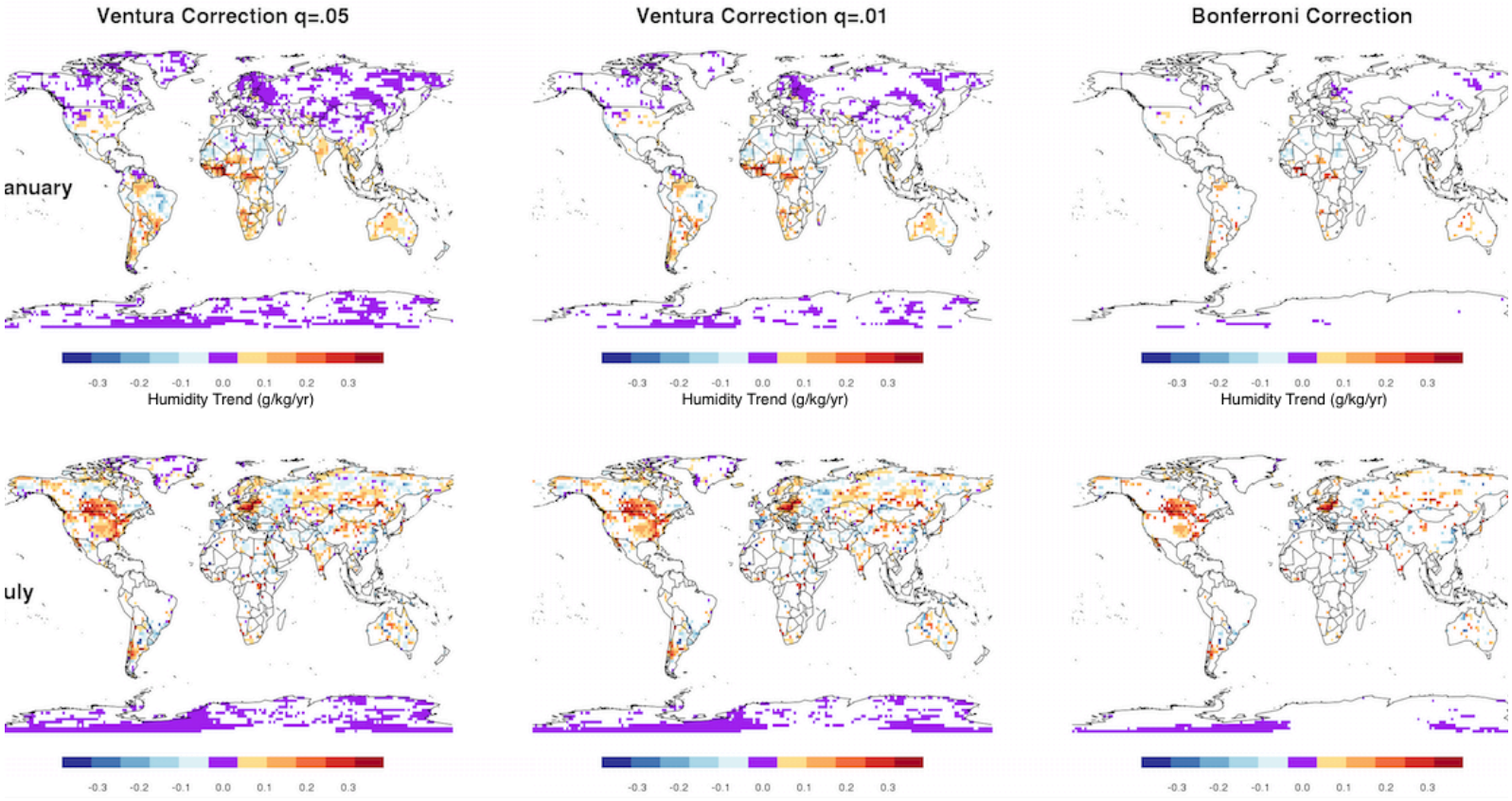

Figure 4. Same as in Figure 3 except with 3 additional correction methods. The Bonferroni method is the most conservative, shown visually by having the most non-significant (white) trends. Note the consistent positive trend in humidity over the eastern U.S. and Canada.

% of Tests Remaining Significant

Method	January Global	July Global	January US & Canada	July US & Canada
Original Trend	n=2438 significant trends out of 5914* tests	n=2724 significant trends out of 4886* tests	n=221 significant trends out of 634 tests	n=394 significant trends out of 599 tests
Benjamini & Hochberg	75%	90%	62%	94%
Benjamini & Yekutieli	37%	64%	33%	78%
Ventura q=.05	99.9%	100%	100%	100%
Ventura q=.01	59%	83%	53%	94%
Bonferroni	14%	36%	19%	58%

Table 1. Percentage of tests that remained significant after applying correction methods. The Ventura method with q=.05 had virtually no effect, while the Bonferroni correction is the most conservative.
*Though the grid is 192x94, only ~33% is land, equaling ~6000 potential tests. Some land pixels had no heat waves over the entire study period, explaining the 5914 vs. 4886 difference.

4.2 U.S. and Canadian humidity trends

The results are surprisingly similar to those when the analysis covered the entire globe. The main difference can be seen in Table 1: the correction methods have less of an impact when restricted to this region than when looking globally.

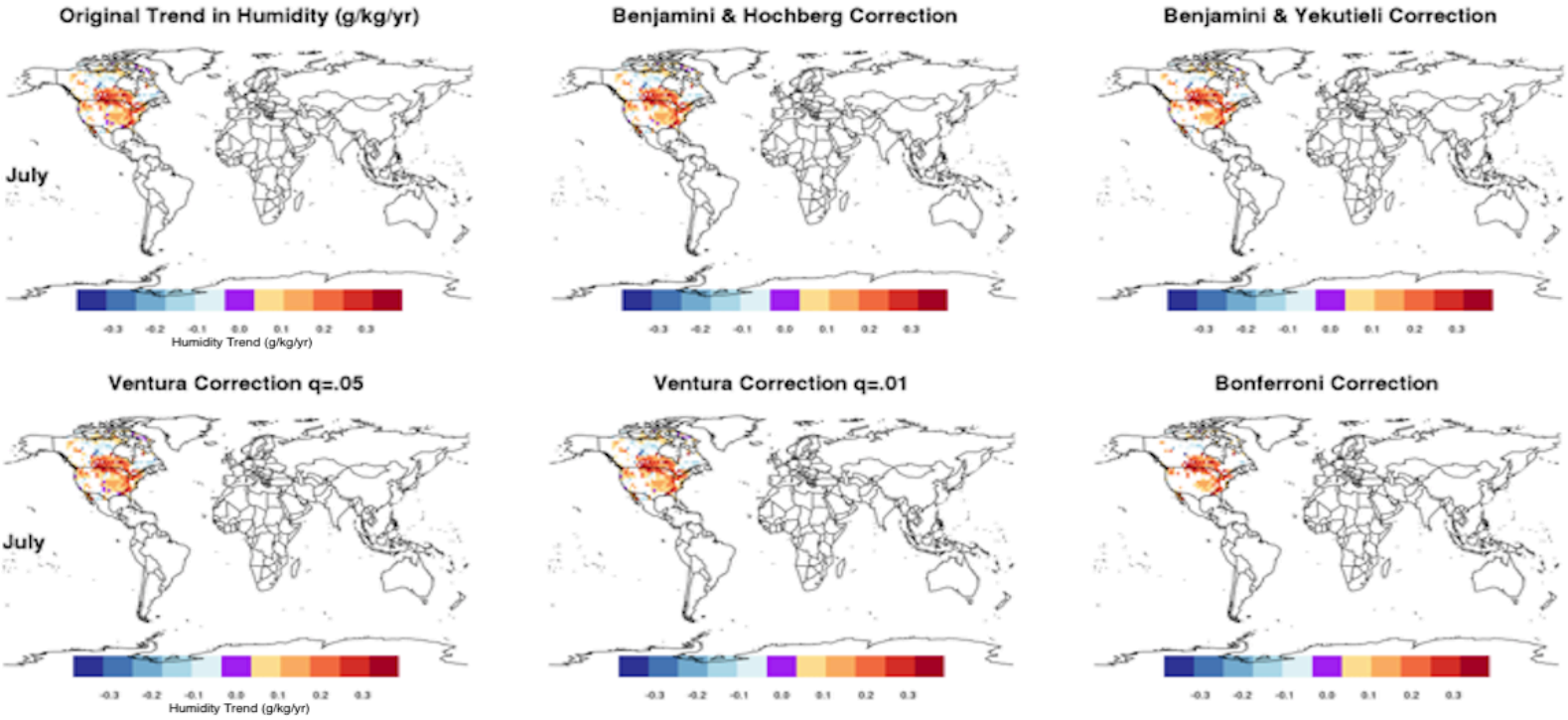

Figure 5. Same as in Figures 3 and 4 except for July only and calculated from the start only over the U.S. and Canada. Note that now since few trends that are significant are close to zero (purple), some of the blue/red regions become white when applying the correction methods, although this is only apparent if one squints very hard.

5. Discussion and Conclusion

Various multiple comparison correction methods were implemented to try and limit the FDR (Type I error) in the calculated humidity trends. The methods in general determine a new cutoff p-value less than the original p-value threshold. This leads to fewer rejections of the null and thus fewer 'significant' humidity trends. Therefore this also mainly impacts the trends whose magnitudes were relatively small, shown in purple in Figures 3-5. With that in mind, visually one generally picks out the greatest magnitude trends, so it is not clear that applying these corrections would necessarily change one's visual conclusions. However, if including these spatial trends in an additional model then these correction methods may become much more important. That being said, there does not appear to be much harm in applying the correction methods and despite the debate it appears to be the 'best practice' so far.

The correction methods in order from most conservative to least were Bonferroni, BY, Ventura with q=.01, BH, and Ventura with q=.05. Although the Ventura method purportedly is preferable for spatial data, it appears to be heavily reliant on choice of q. The default q=.05 used in the original paper lead to no adjustments at all in most scenarios, which suggests the method is flawed, something also found by Wilks (2006). The Bonferroni correction was included mostly to demonstrate the impact of an overly strict correction procedure. The decision between BH and BY is difficult without simulation data or a training/test set evaluation, but subjectively the BH method appears a good compromise between being too conservative and not conservative enough, despite the reduced assumptions of BY corresponding better to the data.

6. Aside

I repeated the entire analysis done here but for rank-based regression instead of least squares regression. Rank-based regression is a nonparametric regression method that among other things is in general less susceptible to outliers. However, the trend value maps and percentage of trends remaining significant were nearly identical to those using least squares regression (not shown, but in the zipfile).

7. References

Benjamini, Y., and Hochberg, Y., 1995, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing: J.R. Statist. Soc., v. 57, no. 1, p. 289–300.

Benjamini, Y., and Yekutieli, D., 2001, The control of the false discovery rate in multiple testing under depencency: The Annals of Statistics, v. 29, no. 4, p. 1165–1188, doi: 10.1214/aos/1013699998.

Della-Marta, P.M., Haylock, M.R., Luterbacher, J., and Wanner, H., 2007, Doubled length of western European summer heat waves since 1880: Journal of Geophysical Research, v. 112, no. D15, p. D15103, doi: 10.1029/2007JD008510.

Frich, P., Alexander, L. V., Della-Marta, P., Gleason, B., Haylock, M., Tank Klein, a. M.G., and Peterson, T., 2002, Observed coherent changes in climatic extremes during the second half of the twentieth century: Climate Research, v. 19, no. 3, p. 193–212, doi: 10.3354/cr019193.

Gasparrini, A., and Armstrong, B., 2011, The Impact of Heat Waves on Mortality: Epidemiology, v. 22, no. 1, p. 68–73, doi: 10.1097/EDE.0b013e3181fdcd99.

Gelman, A., Hill, J., and Yajima, M., 2012, Why We (Usually) Don't Have to Worry About Multiple Comparisons: Journal of Research on Educational Effectiveness, v. 5, no. 2, p. 189–211, doi: 10.1080/19345747.2011.618213.

Habeeb, D., Vargo, J., and Stone, B., 2015, Rising heat wave trends in large US cities: Natural Hazards, v. 76, no. 3, p. 1651–1665, doi: 10.1007/s11069-014-1563-z.

Hochberg, Y., 1988, A sharper bonferroni procedure for multiple tests of significance: Biometrika, v. 75, no. 4, p. 800–802, doi: 10.1093/biomet/75.4.800.

Holm, S., 1979, A Simple Sequentially Rejective Multiple Test Procedure: Scandinavian Journal of Statistics, v. 6, no. 2, p. 65–70, doi: 10.2307/4615733.

Hommel, G., 1988, A stagewise rejective multiple procedure based on a modified Bonferroni: Biometrika, v. 75, no. 2, p. 383–386.

Kanamitsu, M., Ebisuzaki, W., Woollen, J., Yang, S.K., Hnilo, J.J., Fiorino, M., and Potter, G.L., 2002, NCEP-DOE AMIP-II reanalysis (R-2): Bulletin of the American Meteorological Society, v. 83, no. 11, p. 1631–1643+1559, doi: 10.1175/BAMS-83-11-1631.

Pal, J.S., and Eltahir, E.A.B., 2015, Future temperature in southwest Asia projected to exceed a threshold for human adaptability: Nature Climate Change, v. 6, no. 2, p. 197–200, doi: 10.1038/nclimate2833.

Perkins, S.E., 2015, A review on the scientific understanding of heatwaves—Their measurement, driving mechanisms, and changes at the global scale: Atmospheric Research, v. 164-165, p. 242–267, doi: 10.1016/j.atmosres.2015.05.014.

Perkins, S.E., and Alexander, L. V., 2013, On the Measurement of Heat Waves: Journal of Climate, v. 26, no. 13, p. 4500–4517, doi: 10.1175/JCLI-D-12-00383.1.

Perkins, S.E., Alexander, L. V., and Nairn, J.R., 2012, Increasing frequency, intensity and duration of observed global heatwaves and warm spells: Geophysical Research Letters, v. 39, no. 20, p. n/a–n/a, doi: 10.1029/2012GL053361.

Rothman, K.J., 1990, No Adjustments Are Needed for Multiple Comparisons.: Epidemiology, v. 1, no. 1, p. 43–46, doi: 10.1097/00001648-199001000-00010.

Sherwood, S.C., and Huber, M., 2010, An adaptability limit to climate change due to heat stress: Proceedings of the National Academy of Sciences, v. 107, no. 21, p. 9552–9555, doi: 10.1073/pnas.0913352107.

Wilks, D.S., 2006, On “field significance” and the false discovery rate: Journal of Applied Meteorology and Climatology, v. 45, no. 9, p. 1181–1189, doi: 10.1175/JAM2404.1.

Willett, K.M., Dunn, R.J.H., Thorne, P.W., Bell, S., De Podesta, M., Parker, D.E., Jones, P.D., and Williams, C.N., 2014, HadISDH land surface multi-variable humidity and temperature record for climate monitoring: Climate of the Past, v. 10, no. 6, p. 1983–2006, doi: 10.5194/cp-10-1983-2014.