doi:10.1038/nature16467

Supplementary Discussion

Relative influence of larger and smaller EWD impacts. To examine the relative influence of varying disaster impact severity on the overall disaster signal in the composites, we constructed histograms of eventa year differences from control for extreme heat and drought. The distribution of eventa year differences from control spans positive and negative values, with more values falling in the negative (reda shaded areas, Extended Data Fig. 1). In all cases, upwards of 70% of points fall within 20% of respective resampled control means, and fewer than 10% of disasters reflect a deficit of more than 30%. The disaster response signals are thus driven by moderate deficits in production, yield, and area, with a smaller influence of more severe impacts. This finding emphasizes the global agricultural importance of moderately severe disasters.

The influence of sample size. The SEA method as applied here functions under the assumption that if the number of time series composited is sufficiently large, then the signal due to all variables other than the disaster (for instance policy changes or economic shocks) should be distributed evenly positively and negatively relative to control, and therefore disappear after averaging. The composite should thus exhibit only the impact of the disaster. While no definitive guidelines exist for what qualifies as a sufficient sample size for effective compositing, previous applications of SEA have employed samples of about 20[31,32] and as low as 9 constituent time series[33].

Sample sizes by year are tabulated in Extended Data Tables 2 and 4. It should be noted that sample sizes differ across years in a given composite due to varying numbers of coinciding disasters per year removed during compositing to isolate the impacts of specific disaster types (*see* Methods). At some points in the main letter, we present sample sizes averaged across the years for clarity. Respective control composites were resampled using corresponding year-specific sample sizes. While most of our analyses involved much larger samples than those assessed in previous applications, in some cases our sample sizes seemed potentially too small for effective compositing (less than 30, and as small as 14). For cases with small sample sizes, we questioned whether our results may have included type I (persistent noise due to insufficient sample misconstrued as disaster signal) and type II (insufficient sample to isolate disaster signal) errors.

Regarding type I errors, there are two effects of sample size on statistical significance in our methodology, which must be considered together. First, with lower sample size, each disaster composite exhibits more noise relative to signal compared to larger samples, because the probability of non-disaster signals in each time series contributing to the composite canceling out (i.e. being equally positive and negative relative to control) increases with sample size. This effect is visible in the greater variability in the extreme heat and cold composites (n~50) compared to those for flood and drought (n>200) in Figures 1 and 2. Second, at lower sample size, each control composite is more variable, so the distribution of sets controls is wider (visible in the difference in widths of control boxplots between disaster types in Figure 1).

If a deficit signal in the disaster composite is significant, it should dip below most of the 1000 controls (by our criterion, all but 5 of 1000 for two-tailed 99% significance). Otherwise, the signal is not statistically significant (i.e., is within the variability of false-

disaster controls). Since the distribution of controls depends on number of samples in each replicate, the greater variability in disaster composites based on fewer samples corresponds to equivalently wider control distributions (i.e., equivalently more conservative thresholds for significance). We therefore consider this method of significance testing robust to sample size, and are confident that our significant findings in cases with fewer samples do not reflect type I errors.

Type II errors could arise if the sample size is insufficient to isolate a strong mean disaster signal that is differentiable from variability in controls. This may arise due to a saturation effect in which the mean disaster signal becomes increasingly different from control with additional samples as it asymptotically approaches the final estimate. The existence of a saturation effect would call into question, for instance, whether wheat and rice genuinely respond less to extreme heat, or if the sample is simply too small for SEA to isolate an effect in rice and wheat (Figure 4).

To examine whether such a saturation effect exists, we performed an illustrative resampling experiment in which we computed 200 pseudo-mean disaster impacts on 16-cereal aggregate production using random sub-samples of the full disaster sets of size (1, 2, … , n). This procedure enabled us to visualize the convergence of and variability in 200 possible paths to our actual mean disaster impact at the full sample size (Extended Data Figure 2). We used the set of 200 estimates at n-1 to estimate the 95% confidence intervals of our full-sample impact estimates[34]. At low sample sizes, the impact of extreme values on the mean is greater, resulting in high variability among the different estimates. With increasing samples, the variability of estimates reaches an inflection point after which incremental samples result in only small incremental decreases in the variability of estimates. With increasing samples below this point, the

majority of noise is reduced by compositing. The fact that our actual sample sizes are far beyond this zone lends us confidence that our sample sizes are sufficient.

If failures to reject the null hypothesis in cases with small samples were due to insufficient sample size, then there would exist a bias towards under-estimating the disaster signal relative to final estimate when resampling with smaller sample size ($N<n$). In other words, if a saturation effect exists, then the pseudo-estimates at $N<n$ should approach the final estimate at $N=n$ from the positive side. In that event, the disaster signal may become significant with additional samples. Since the pseudo-estimates are roughly evenly distributed on either side of the final estimate (grey dotted line, Extended Data Figure 2), we deduce that no saturation effect is responsible for the lack of significant findings for these cases, and that our lack of findings in these cases are not reflective of a type II error. The fact that significant signal was detected in some cases with equivalently small samples provides us further confidence in our findings.

Comparative statistics and assumptions. For the individual crop, regional, and earlier-versus-later droughts analyses, we assessed the significance of differences in disaster impacts between groups by applying Kruskal-Wallis one-way non-parametric analysis of variance to the sets of individual disaster responses in each grouping. We made this choice because, after applying the Anderson-Darling test for normality to the data, we rejected the null hypothesis of normality at the 5% significance level for all groups, and therefore could not use parametric statistics which assume a normal distribution.

Kruskal-Wallis analysis of variance tests the significance of differences between group distributions (without the assumption of normality), which includes differences in both mean and variance. In comparing disaster impacts among regions, crops, and times, differences in mean were more salient for our study than differences in variance.

An initial application Levene's Absolute test for equality of variances on the data revealed unequal variance for many groupings that we wished to compare (at 5% significance). Although inequality of variance does not violate the assumptions of the Kruskal-Wallis test, it complicates the attribution of statistically significant differences to mean versus variance.

Applying a quadratic transformation to the entire set of disaster-year responses resolved the issue of unequal variance in most cases (Extended Data Table 6), and did not in any way affect the Kruskal-Wallis results. A re-application of the Anderson-Darling test revealed that the transformed data still deviated substantially from normality. We therefore proceeded with quadratically-transformed data and the Kruskal-Wallis test in assessing our comparisons for statistical significance. Where differences between groups were significant but variances unequal, we have stated that we cannot conclude whether the significance is due to differing mean or variance. This situation arises most notably for individual crop yield for extreme heat (Figure 4d). In this case, while the Kruskal-Wallis results fail to conclusively differentiate the mean signals among crops, maize is still the only crop with significant yield impacts.

Comparison to previous studies. We cannot compare our estimated yield deficits from extreme weather events to previous studies of the impact of seasonal mean climate trends over the same period[20] because we quantify mean per-disaster sensitivities and not impacts due to climate trends. To compare our yield loss estimates based on reported disasters to previous studies based growing season weather anomalies, it was necessary to establish whether the reported disasters correspond to significant seasonal weather anomalies. We therefore repeated the compositing procedure using nationally-averaged JJA (DJF in Southern Hemisphere countries) mean temperatures

and total precipitation from CRU TS 3.2[35] for our samples of extreme heat and drought disaster over 1961-2010, and multiplied the percent anomalies by the global averages. Drought did not correspond to any precipitation anomalies and showed only slight temperature anomalies of 0.15ºC (Extended Data Figure 3b-c). We conclude that the drought impacts observed here are therefore independent from those estimated in previous studies based on seasonal mean anomalies. Meanwhile, the extreme heat disasters corresponded to a significant mean temperature anomaly of +1.2ºC, which implies a yield sensitivity of 6-7% per 1ºC (Extended Data Figure 3a). This value falls within the range of sensitivity estimates in Lobell and Field (2007)[36], which found 5-9% yield reductions per 1ºC increase in seasonal temperature for a number of crops. However, we cannot draw any deeper conclusions by comparing our numbers due to the following methodological differences:

1) Comparing the two estimates for yield impacts at ~1ºC assumes that the years in each sample feature comparable extremes. But since the relative contribution of moderate versus extreme warm days to the mean anomaly is unknown, the two samples cannot be assumed to reflect similarly extreme weather perturbations to the crops.

2) Estimated sensitivities in Lobell and Field (2007) are based on linear regressions covering a temperature domain of about -1 to 1ºC in which the vast majority of points fall below 1ºC. The sensitivities based on these linear regressions cannot be assumed to be powerful at and above the maximum temperature in their domain, and therefore may not reflect the level of extremes that we examine.

3) Our estimates are based on globally-averaged national weather and crop data, while those in Lobell and Field (2007) are based on data averaged across cropping regions that transcend borders[36]. Because the estimates are made over differing spatial

scales, they may not reflect variation in yield and weather similarly. The disasters in our study do not include sub-national spatial data, so we cannot at present solve this discrepancy by spatially disaggregating.

Number of disasters per year. The EM-DAT database is based on a compilation of disaster reports gathered from various organizations including United Nations agencies, governments, and the International Federation of Red Cross and Red Crescent Societies. A time-series of reported disasters per year (Extended Data Figure 4) exhibits an upward trend. This is likely a result primarily of an improvement in the completeness of disaster reporting over time, but the trend may also partially reflect greater incidence of disasters in recent decades. Assuming that incomplete disaster reporting in earlier decades was independent of disaster impact severity, our estimates of cereal crop sensitivities to EWDs are robust to this trend because we quantify them on a per-disaster basis. But we likely underestimate the total global cereal production lost to these disasters because we aggregated losses over an incomplete set of disasters.

References

31. Adams J. B., M. E. Mann, and C. M. Ammann. 2003. Proxy evidence for an El Niño-like response to volcanic forcing. *Nature* 426: 274-278.

32. Lühr, H., M. Rother, T. Iyemori, T. L. Hansen, and R. P. Lepping. 1998. Superposed epoch analysis applied to large-amplitude travelling convection vortices. *Annales Geophysicae* 16(7): 743-753.

33. Mass, C. F., and D. A. Portman. 1989. Major volcanic eruptions and climate: A critical evaluation. *J. Climate* 2: 566-593.

34. Efron, B., and C. Stein. 1981. The jackknife estimate of variance. *The Annals of Statistics*: 586-596.

35. Harris, I., P. D. Jones, T. J. Osborn, and D. H. Lister. 2014. Updated high-resolution grids of monthly climatic observations – the CRU TS3.10 Dataset. *Int. J. Climatol*. 34: 623–642.

36. Lobell, D.B, and C. B. Field. 2007. Global scale climate-crop yield relationships and the impact of recent warming. *Environmental Research Letters* 2(1).