Methods section

To get a sense of how stressful the weather during this field study was, relative to long-term summer weather data, the 2009 KDTW *(network figure)* observations were compared to the past 30 years of weather. Apparent temperatures (Steadman 1979a, 1979b, 1984) were used, rather than conventional air temperatures, because the Great Lakes Region often has high levels of water vapor and heat stress is dependent on water vapor content. From 1979-2009 with the KDTW observing station's data, the apparent temperature each hour was calculated, and then the maximum (minimum) temperature was taken as the highest (lowest) apparent temperature observed between one and seven pm (am) (EDT or GMT-4). We used 1979-2008 as a base period and a sub sampling method within that base period of a 19-day window centered around the Julian date in question. The percentiles were determined empirically, and thus do not make any assumptions on the underlying distribution. Subsequently, the percentile for each day during the 2009 field study was determined, for both daily extremes, w.r.t. 570 (19x30) observations of apparent temperatures experienced in Detroit around that Julian date. We then plotted the time series of both daily high and daily low apparent temperature percentiles and calculated the 110-day mean (110DM) percentiles.

As a starting point, the basic statistics of the magnitude of spatial variability across the region were calculated. The magnitude of spatial variability was taken - each day and for both daily extremes - as the *range* in the anomalies, across all sensors, and is referred to as the *UHI magnitude*. This was done for each of the 110 daily highs and lows of the field experiment. First the 110-day mean, standard deviation and maximum value UHI magnitudes were calculated, for both daily extremes. A *Student's T-test* was then used to determine statistical significance (at the five percent significance level) in the summer average UHI magnitude. Lastly the histograms for the UHI magnitude, in both daily extremes, were calculated and graphed.

Further investigation of the temporally averaged observed temperature structure was initially done by qualitative evaluation of the spatial pattern with respect to the major features of Detroit (e.g. water bodies, city layout, etc.). To accomplish this we plotted both (e.g. max & mins) of the 110DM anomalies on maps of the metropolitan region. Additionally, investigation into the stability (e.g. consistency) of that spatial pattern was by undertaken by comparing the range in the 110DM anomalies to the 110DM UHI magnitude. The former requires no stability and the later is reduced by instability, and thus we compared the ratio of the two values in both daily highs and lows. In order to describe the temporal *signal-to-noise* characteristics we graphed the stations' 110DM anomalies with their corresponding standard deviations. Lastly, a bootstrapping procedure with 2,000 resamples was used to better estimate the true mean anomalies and the confidence in those means. This was done by resampling (with replacement) from the 110-day sample, 2,000 times, and then determining the mean each time. The mean of these means is a better estimate of the true mean, and the variability can be quantified by either the standard deviation, or empirically, by taking the 2.5th and 97.5th percentiles (of the calculated means).

The importance of the UHI phenomenon in Detroit, to public health officials and urban planners, depends on whether it occurs during dangerously hot summertime weather. Because the definition of "dangerous" is not clear we choose to evaluate the intersection of the UHI magnitude with days labeled dangerous by multiple methods. Generally, these methods characterize days as oppressive if the weather conditions likely result in high heat stress for the general public, and if the conditions are sustained it is termed a heat event. Lastly, any trends the UHI magnitude had with increasing apparent temperature percentiles was investigated. These methods are more fully described below.

Initially we sought to use heat-health warning system (HHWS) methods to determine which days were especially stress w.r.t. heat. The first method was the air-mass-based Spatial Synoptic Classification 2 (SSC2) (Sheridan 2002; Kalkstein and Sheridan 2003) method. In this method each day is determined to be of a specific "air mass" type, and in Detroit "Dry Tropical", "Moist Tropical +" and "Moist Tropical ++" are the dangerous air masses. Calendar data provided by Dr. Scott Sheridan's website (Sheridan 2010) was used for this purpose. The mean UHI magnitude was calculated as a function of all air masses observed, rather than just the oppressive ones. The second method was the taken as from NWS heat advisory. Daily highs and lows were selected that occurred during period of advisories as indicated from the NWS's *Non-Precipitation Warnings, Watches, Advisories* bulletins (NCDC HDSS Access System reference). The third method was taken from a study in which Easterling et al. (2007) defined a *Heat Wave*. The definition requires three days with the daily high exceeding the 80th percentile of daily highs and subsequent daily low exceeding the 80th percentile of lows. Once such heat events were determined during the study period, then the mean UHI magnitude was calculated during those periods. The percentiles used for this were the same as those calculated previously using the apparent temperature from KDTW. If no heat events existed according to that definition, then the individual days indicated as oppressive were analyzed as a group. With all three definitions, if possible a Student's T-test was employed to determine whether the observed mean UHI magnitudes' were statistically significant.

A more direct way to determine UHI magnitude relevancy during hot weather was preformed empirically. The previously determined daily high and low apparent temperature percentiles, observed at KDTW, were again used here. Ordinary least squares regression between the observed UHI magnitude and apparent temperature percentile values was used to determine the magnitude of trends and their significance. The 95 percent confidence intervals of the regression determined slopes were calculated as well as the y-intercepts at the extrapolated 100th percentiles. Subsequently, relevancy of the UHI magnitude as a function of heat stress was speculated upon.

The dynamic nature of the UHI magnitude led to investigation into whether the variability could be explained, or predicted, by the weather conditions. These weather variables are treated as proxies for physical processes that create both the daytime and nighttime UHI phenomena (e.g. cloud cover as an indicator of daytime (nighttime) radiative heating (cooling) rates). The relationship between each day's UHI magnitude value and the regionally observed wind speed and cloud-cover values was undertaken. The daily low UHI magnitude was evaluated against three variables: previous afternoon average cloud-cover percentage, morning average cloud-cover percentage and morning average wind speed. The daily high UHI magnitude was only tested against two variables: average afternoon cloud-cover percentage and average afternoon wind speed. The observations were spatially averaged across three surrounding airports (KVLL, KDET and KDTW) (*Network figure*); and temporally averaged for the morning between four and eight AM, and the afternoon between two and six PM (all EDT, or GMT-4). Once the relationship with the airport observations was investigated we repeated the investigation with corresponding reanalysis data. This was done with regards to the ability to forecast Detroit's UHI magnitude. The North American Regional Reanalysis data was used and similarly spatially averaged (e.g. grid points that the three airports were located within) and temporally averaged (e.g. only the times within the time ranges previously used). The wind speed values, for both afternoon and nighttime, were calculated from NARR's *3-hourly value of U-Component at 10-meters* and the analogous *V-Component* variable. To represent cloud cover the *3-hourly forecast of Total Cloud Cover of the entire atmosphere* used for the overnight periods and the *3-hourly average Downward Shortwave Radiation Flux at the surface* for the afternoon periods.

For each predictor variable the *Spearman Rank* (SR) correlation coefficient was calculated to the respective predictand (daily high and low). Then for both predictands, stepwise multiple regression was performed using the variables deemed appropriate by the backwards elimination method. Evaluation of the regression coefficients, determined from scaled inputs, was done in order to determine each predictor's relative contributions to the models. To determine the goodness of those regression models, statistics such as p values (p), root mean squared error (RMSE) and amount of variance explained (R^2) were calculated and compared to both the mean and maximum UHI magnitude. Multiple regression models that passed initial scrutiny were further tested for their functionality by evaluating their predictive ability in a cross-validation manner. This was done by first removing every fifth day from the 109-night sample (the first day was excluded because we did not have a previous day's afternoon cloud cover), building the regression model with the days not eliminated, and then using that model to predict the UHI magnitude values of the removed days. We calculated the RMSE and average absolute difference, and then compared those to the maximum and mean UHI magnitude (of the evaluation sample (n=22)). Additionally, the confidence in the determined regression coefficients was evaluated using a bootstrapping technique to evaluate regression coefficients. This was accomplished by creating

2000 bootstrap samples, by sampling with replacement from the 109-sample, and then determining the regression coefficients each time. The 2000-sample mean regression coefficients are a better estimate of the true coefficients, and were calculated. Additionally, the confidence can be inferred by the standard deviations and they were also calculated.

Next, investigation into the relationship between spatial attributes and observed temporally averaged anomalies was undertaken. While these relationships were built and then assessed using the observing stations, generally, spatial attributes are continuous in nature and thus can help infer in the space between the observing stations. We choose three spatial attributes that are referred to as "distance to city center" (D2CC), "distance to water" (D2H2O) and "percent impervious surface" (PIS). The values of these variables were calculated at each monitoring location with the help of ArcGIS software (ESRI Corporation, Redlands/CA). Then the SR correlation coefficients between these different spatial variable values were calculated. Then investigation into whether their relationships with the 110DM anomalies was undertaken and the spatial variables are more fully described below.

The first variable, PIS is an indicator of the built environment (Oke 1973; Arnold and Gibbons 1996). Impervious surfaces are the hard constructed surfaces that cover buildings, roadways, parking lots, etc. Areas surrounded by a greater PIS may be more likely to store heat, and then release that heat overnight. The imperviousness spatial data used was from the National Land Cover Data 2001 Imperviousness dataset, which is derived from Landsat imagery taken in 2001 (U.S. Geological Survey 2008). Maps were constructed with both the sites of the monitoring sites and a superimposed data layer of per-pixel (30 m resolution) estimates of imperviousness. Using ArcGIS software, the percent of surrounding surface indicated as impervious was calculated for each station. Imperviousness does not take into account the 3-dimensional geometrical factors or distance from city center. Initially investigated were the relationships between the temporally averaged mean anomalies, both highs and lows, for each station and the PIS values within various circular radii from 0.15 to 3.0 km. SR correlation coefficients were calculated between each radius' PIS values (at each station) and the 110DM anomalies. This was done in order to determine the strongest relationship between local percent impervious values and both daily extremes. Subsequent assessment of the significance of the strongest relationships was performed.

The second spatial variable, D2H2O was a logical driving mechanism of temperature variability. Higher thermal inertial of a water body could dampen the diurnal cycle and then would impact both daily extremes of near-shore locations. Specifically with regards to the daily highs, both smaller scale 'lake breezes' and larger scale cold-air advection could cool near-shore regions. With regards to the daily lows, high lower-level atmospheric water vapor content at near-shore locals could decrease long wave radiative cooling rates at the surface and subsequently affect shelter height temperatures. To investigate this driving mechanism, the correlations between the straight-line distance to a sizeable water body at each station and the observed study-duration mean anomalies were calculated. Relevant water bodies were taken here as the Detroit River, Lake St. Clair and Lake Erie and the distance was drawn from the land-water boundary nearest to each location.

Next, distance from city center was examined as a driving mechanism of the temperature variability. The general idea of an UHI phenomenon consists of concentric isotherms surrounding a city (coolest on the outside). Daily highs might be higher at the city center because of increased anthropogenic heat flux and roughness length nearer the city center. Roughness length retards the rate of heat loss (upwards) within the urban canopy layer. Additionally, latent heat flux and short-wave albedo (short-wave radiation trapping) should be lowest in the city's interior. However, one can make the counter-case in Detroit, with the water bodies' proximity to downtown and declining downtown population. Alternatively, the relationship between the daily lows and distance from city center is uncomplicated, considering the anthropogenic heat flux, urban-canyon effect (e.g. long-wave radiation trapping) and volumetric heat capacity are expected to be largest at the city center.

We then evaluated the ability to predict the 110DM anomalies from the aforementioned spatial attributes. To systematically determine which variables were most appropriate the backwards elimination stepwise regression method was employed with all three variables as predictors. In order to evaluate the determined regression equations the RMSE and percent of variance explained (R^2) were calculated and for significance the full-model p value (p) were calculated. A brief evaluation within each equation (e.g. strength of the standardized correlation coefficients) and between equations (e.g. predicting the daily highs and lows) was undertaken. To test the utility of the regression equations that were deemed significant, we predicted each station's temporally averaged temperature anomalies using the other 29 stations, in a leave-one-out cross-validation manner. The goodness was evaluated by calculating the RMSE and average absolute error and compared them to the range in the anomalies being predicted. Lastly the confidence in the empirically determined regression coefficients was evaluated with a bootstrapping technique where 2,000 resamples (with replacement) of the stations values were made and the regression coefficients calculated for each resample. Then the mean, of those 2,000 means, regression coefficient was calculated as well the standard deviations.

To get a more robust picture of the UHI characteristics in Detroit a brief exploration was done into the observations taken only during days where a large spatial variability in temperatures is expected. The bootstrapped derived regression equation between the weather conditions and the UHI magnitude was used to sub select days which we would expect a large UHI magnitude. We decided to select days in which the equation predicted UHI magnitude was larger than the median predicted UHI magnitude; thus the new sample size was 55 days. For comparative purposes, we recalculated the signal-to-noise ratio. Done here by calculating the mean anomaly and confidence intervals per station, again using the bootstrapping method (to minimize effect of small sample size) with 2,000 resamples. Also we calculated the mean UHI magnitude, the range in mean anomalies and then the ratio between the two. Lastly, the spatial distribution of mean anomalies was again plotted atop a map of Detroit.

Evaluation was then undertaken of the ability to predict the sub selected sample average anomalies from the spatial variables. For simplicity, this was executed only for the daily lows. The backwards elimination stepwise regression method was used and indicated that all three variables as predictors to predict the sub selected average anomalies was appropriate. Again the RMSE and percent of variance explained (R^2) were calculated, and for significance the full-model p value (p) were calculated for evaluative purposes. Comparing the coefficients calculated using normalized predictor variables allowed for quantification of which variables were most influential. To further test the ability of the spatial attributes to predict the average anomalies we estimated each station's anomaly in the leave-one-out cross-validation manner previously described; then calculated the RMSE and average absolute error in predicting the true values. Lastly, we used the same bootstrapping method as described for the 110DM anomalies to estimate the true regression coefficients and subsequently assessed our confidence in the regression coefficients calculated from the observed sample.