INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. **31**: 200–217 (2011)
Published online 15 April 2010 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/joc.2141

RMetS
Royal Meteorological Society

A systematic review and scientific critique of methodology in modern urban heat island literature

I. D. Stewart*
Department of Geography, University of British Columbia, Vancouver, BC Canada

ABSTRACT: In the modern era of urban climatology, much emphasis has been placed on observing and documenting heat island magnitudes in cities around the world. Urban climate literature consequently boasts a remarkable accumulation of observational heat island studies. Through time, however, methodologists have raised concerns about the authenticity of these studies, especially regarding the measurement, definition and reporting of heat island magnitudes. This paper substantiates these concerns through a systematic review and scientific critique of heat island literature from the period 1950–2007. The review uses nine criteria of experimental design and communication to critically assess methodological quality in a sample of 190 heat island studies. Results of this assessment are discouraging: the mean quality score of the sample is just 50 percent, and nearly half of all urban heat island magnitudes reported in the sample are judged to be scientifically indefensible. Two areas of universal weakness in the literature sample are *controlled measurement* and *openness of method*: one-half of the sample studies fail to sufficiently control the confounding effects of weather, relief or time on reported 'urban' heat island magnitudes, and three-quarters fail to communicate basic metadata regarding instrumentation and field site characteristics. A large proportion of observational heat island literature is therefore compromised by poor scientific practice. This paper concludes with recommendations for improving method and communication in heat island studies through better scrutiny of findings and more rigorous reporting of primary research.

KEY WORDS urban climatology; heat island magnitude; scientific method; systematic review; critical analysis

Received 23 October 2009; Revised 1 March 2010; Accepted 4 March 2010

1. Introduction

Observations of the urban heat island (UHI) effect have a long and well-documented history in climate literature. In 1833, the first scientific observations were documented by Luke Howard, whose temperature analysis in and around London, England, portrayed a city distinctly warmer than its countryside. Howard's observations were motivated in part by the fact that meteorology was 'less trodden' than other disciplines, and thus it was lacking the 'regular and consistent form of a science'. In the two centuries that have passed since Howard's temperature observations, heat island studies have been published in hundreds of cities worldwide, including almost every major city in Europe, North America and East Asia. These studies and their estimates of UHI magnitude are unrivalled in their contributions to urban climatology, and comprise a literature of great historical and geographical interest.

The overwhelming size of this literature – produced by a relatively small and diverse group of scientists – is reason enough, however, to question the authenticity with which heat island observations have been gathered and reported through history. To what extent does this literature serve the aims of science? By what judgement does is it constitute 'sophisticated' observations? Can its measurements be trusted? One can quickly surmise from standard reviews of heat island literature that a response to these questions is not obvious, but that evidence on which to hypothesise is plentiful. Modern heat island investigators such as Parry (1956), Chandler (1962, 1970) and Bohm and Gabl (1978), for example, alluded to problems of methodology decades ago. In recent years, discussion around these same problems has been open and direct (e.g. Oke, 2006, 2009; Stewart, 2007).

A formal assessment of modern heat island literature is now both timely and necessary. This paper reflects on the scope and status of that literature through a systematic review and scientific critique of its primary observational studies. One hundred and ninety sample studies from 1950 to 2007 were appraised for their scientific quality based on criteria of experimental design and communication. Although the systematic review finds certain strengths in the literature, these are largely overshadowed by universal weaknesses in definition, measurement and communication. This paper closes with specific recommendations for improving methodological quality in UHI literature.

* Correspondence to: I. D. Stewart, Department of Geography, University of British Columbia, 1984 West Mall, Vancouver, BC Canada, V6T 1Z2. E-mail: stewarti@interchange.ubc.ca

2. Methods

The traditional and most accessible approach to literature assessment is the standard review, the purpose of which is to describe current knowledge on a topic and to explain recent research findings. Urban climatologists have invested heavily in this tradition, with many reviews examining UHI literature and its rapid growth through the twentieth century (e.g. Kratzer, 1937; Brooks, 1952; Peterson, 1973; Oke, 1979; Landsberg, 1981; Nakagawa, 1996; Arnfield, 2003; Roth, 2007). Standard reviews, however, are seldom critical, they rarely engage the quality of the original studies, and generate little if any new knowledge. Yet, despite these limitations, literature reviews are the most widely cited papers in science (Cooper and Hedges, 1994).

Systematic review differs from *standard* review in that it integrates a body of literature by methodically extracting data from a representative sample of primary studies (Hunt, 1997). The extracted data are then combined into a single 'super study' with quantitative and decisive findings. Reconciling methods with output, systematic review gives coherent, scientific explanations for disorderly and fragmented results in the primary literature. Overarching patterns or problems that are not normally discernible among individual studies then begin to emerge. Systematic review is therefore well suited to topics supporting a substantial volume of accumulated studies.

A systematic review follows four crucial steps that conform to the review protocols of a traditional meta-analysis: (1) the population, or 'universe', of studies about which the review aims to generalise is defined by strict eligibility criteria; (2) a representative sample of that universe is retrieved from the literature through a logical search strategy; (3) essential information from each eligible item is extracted, coded and combined into statistical outcome measures; and (4) the methods, results and theoretical implications of the analysis are reported and discussed. Systematic review by this design is as much a scientific enterprise as the primary research it evaluates.

2.1. Defining the universe of studies

The universe of studies is the complete body of literature about which a review aims to generalise. This review generalises the methodological quality of ground-based observational heat island studies and their estimates of canopy-layer UHI magnitude. A universe of this description includes thousands of studies and extends well beyond the capacity of any single structured review. Strict eligibility criteria are therefore necessary to reduce the study universe to a workable size for evaluation, and to keep the sample coterminous with the universe of reality it is said to represent. The search for a representative and homogeneous sample of studies is a crucial first step of systematic review, and its importance to the external validity of the review cannot be overstated.

2.1.1. Eligibility criteria

Selecting a representative study sample for review and evaluation is a multi-step process (Figure 1). All studies included in the review were screened by three eligibility criteria. Studies that successfully met each of these criteria were declared eligible for further assessment and retained in the sample. Studies failing one or more of the criteria were immediately disqualified from the sample.

(1) Characterisation of the UHI effect

The first eligibility criterion targets canopy-layer, ground-based observational UHI studies of local-meso time and space scales. All studies incorporating stationary or mobile temperature surveys spanning one or several neighbouring urban settlements for the purpose of observation, description, or explanation of the nocturnal UHI effect were successful in meeting the first eligibility criterion. Local-meso scales were confined to horizontal distances of 10^2–10^4 m, and to time periods of days, months or years. The first eligibility criterion disqualifies all investigations defining heat islands by larger or smaller scale sets, or by alternative sampling methods or sensing media. Immediately rejected from the sample were studies of boundary-layer heat islands, remotely sensed heat islands, surface or subsurface heat islands, daytime heat islands, and non-urban heat islands.

(2) Principal aims

The second eligibility criterion targets all studies aiming to quantify UHI magnitude, or intensity, in a specified city, town, village, or other local-scale settlement. This aim invokes empirical measurement of an air temperature differential across city and country, urban and rural, or otherwise built and non-built landscapes. If the candidate study had no intent of quantifying UHI magnitude for a particular settlement, or if this intent was not its principal aim, it was withdrawn from the sample.

(3) Date and source of print or publication

The third eligibility criterion restricts the review to a time period in which no major theoretical or methodological shifts or revolutions changed the field of heat island investigation or its experimental ideals. The chosen period for this review is 1950–2007. This period captures the beginning of the modern era in urban climatology – which is generally ascribed to Sundborg's 1951 classic heat island study of Uppsala, Sweden (Oke, 1995) – as well as first usage of the term "urban heat island" in English-language literature. It also captures the majority of published heat island studies worldwide, including those of tropical and developing regions. All studies printed or published between 1950 and 2007 passed the time filter of the third eligibility criterion. The third criterion further restricts the study sample to the original, or 'primary', works of scholars and researchers.

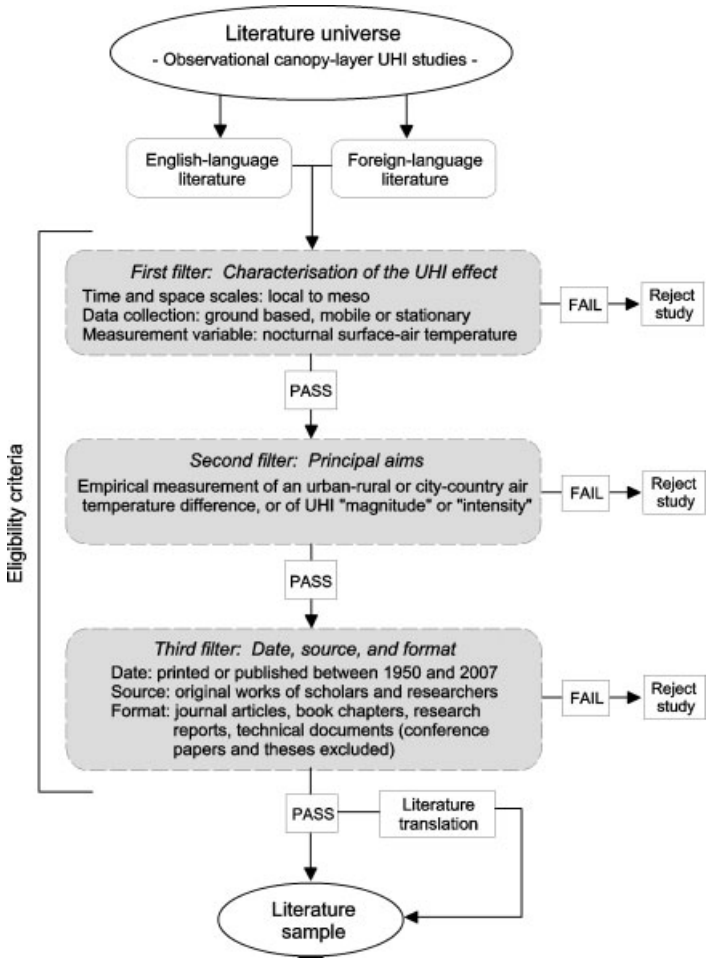

Figure 1. Flow diagram illustrating the selection of literature for review and evaluation.

Editorials, surveys and standard reviews of UHI literature were therefore excluded from the sample. Large quantities of primary research can be found in 'fugitive' literature, which by definition is not widely distributed or indexed and for that reason difficult to locate. Unpublished theses, dissertations, manuscripts, conference papers and newsletters are examples of fugitive literature that exist in large quantities and that were excluded from the sample. Other fugitive sources, such as government reports or institute papers, were included only if the work was original and met all remaining eligibility criteria. Finally, multiple papers by the same author and of the same study area were eligible only if the heat island magnitudes in those papers were derived from different time periods or data-collection methods. Duplicate papers appearing in different sources or languages were immediately disqualified from the sample.

2.2. Sourcing and retrieving the primary literature

The study sample, as defined by the preceding eligibility criteria, was sourced primarily through online and print-accessible abstracts, article indexes and bibliographic databases, both public and private. Additional references were obtained through 'ancestry' searching, which involved manually retrieving citations from bibliographies and reference lists of books, serials, conference proceedings, literature reviews, articles and so on. Expert consultation was also an important link to undiscovered literature. Personal communication with conference and workshop participants uncovered many non-circulating works and historical pieces not indexed in public or private databases, and that usefully expanded the study sample.

The search for a study sample followed a logical strategy to ensure that little time was wasted on irrelevant or irretrievable citations. Hundreds of bibliographic references and article summaries relevant to the study universe were screened for eligibility. Based on title and summary content alone, a large proportion of these were disqualified from the sample for failing one or more of the eligibility criteria. Many studies that met all eligibility criteria were also disqualified, but instead for logistical problems with document recall.

2.2.1. Foreign-language literature

The inclusion of foreign-language literature in a systematic review is vital to its external validity. In keeping the literature sample to a manageable size and averting lengthy and fruitless searches, only a small number of foreign-language studies were included. The candidate list of foreign-language citations was restricted to

major languages of international scientific communication, which include German, French, Russian, Chinese, Japanese and Spanish (Large, 1983). The process of selecting and screening foreign-language citations was identical to that of the English literature (Figure 1). All eligible papers were translated in ad hoc fashion, meaning that only specific material was selected for translation depending on the English-language content of the paper. Foreign-language papers with English abstracts and figure captions required only partial translation, whereas papers with no English content required more comprehensive translation. Full translation was rarely needed of any paper.

Translators were generally non-experts in urban climatology. They were given standardised abstraction forms for retrieving important details from each heat island paper, and were advised to locate and translate verbatim those excerpts responding directly to the critical questions contained in the abstraction forms. These forms ensured that all translators followed an identical abstraction protocol. Finally, translators were warned not to make subjective inferences from vague or missing content of a paper, but instead to convey its factual content as accurately as possible.

2.3. Evaluating the primary literature

The following scientific criteria were developed for the purpose of assessing methodological quality in the heat island literature sample:

- Operational test and conceptual model are aligned;
- Operational definitions are explicitly stated;
- Instrument specifications are explicitly stated;
- Site metadata are appropriately detailed;
- Field sites are representative of the local-scale surroundings;
- Number of replicate observations is sufficiently large;
- Weather effects are passively controlled;
- Surface effects are passively controlled;
- Temperatures are measured synchronously.

These criteria were conceived from (1) well-known methodological and conceptual frameworks in urban climatology (e.g. Landsberg, 1970; Oke, 1976; Lowry, 1977; Goldreich, 1984; Wanner and Filliger, 1989; Szymanowski, 2005); (2) World Meteorological Organization (WMO) guidelines for meteorological observation (e.g. WMO, 1983; Oke, 2004) and (3) classical interpretations of scientific method (e.g. Hempel, 1966; Valiela, 2001). Included in (3) are the hallmark features of science: the *problem statement,* consisting of a conceptual model, operational definitions, and research hypotheses; and *systematic measurement,* consisting of a defined study area and controlled and repeated observations.

The primary studies were assigned a 'pass', 'fail' or 'unknown' grade for each scientific criterion. Any information that was needed to respond positively or negatively to a particular criterion, but that was not available in a report, was termed 'missing data'. Missing data include all feature of experimental design relating to a study's definitions, assumptions, procedures and outcomes. The grading of each primary study by the scientific criteria was based only on evidence contained in its original source document. No supplementary information to favour the decision process was retrieved from external sources, such as the authors themselves or other publications. In this way, poor communication is tantamount to methodological weakness: writing accurate and detailed reports is as much a part of the scientific process as observation itself. Each study's success with the scientific criteria is therefore balanced on sound methodology and effective communication.

2.3.1. *Scientific criteria*

(1) The operational test of the investigation is aligned with the conceptual model of a canopy-layer UHI. The test for this model invokes air temperature measurement below roof level in urban environments, and in the turbulent surface layer of rural environments.

Having the stated or understood aim of measuring UHI magnitude or intensity in the canopy-layer, each study must invoke a suitable test of these concepts. The operational test required of the canopy-layer heat island model is surface-air temperature measurement in urban and rural, city and country, or otherwise built and non-built environments. This model is implicit in Howard's (1833) historical analysis of London's heat island, but is developed and systematised more formally by Oke (1976, 1982, 1988) in modern literature. Studies that fail to measure air temperature at approximately shelter height (1–2 m agl), or at least below roof level, and at field sites broadly defined as urban and rural, are poorly aligned with their conceptual model. These studies met Criterion 1 unsuccessfully. If sufficient detail of instrument height was not found in a report, or could not be inferred from its text, tables or figures, Criterion 1 was graded 'unknown'.

(2) Operational definitions of UHI magnitude or intensity are explicitly stated in the report, or made implicit through its discussion or presentation of data. Operational definitions reveal the measurement variables and field sites used to quantify UHI magnitude.

Operational definitions translate concepts into procedures. Investigators must therefore contrive and communicate appropriate ad hoc procedures of their own to quantify the magnitude of a canopy-layer UHI. Criterion 2 requires two conditions of an operational definition: it must stipulate (1) the location and number of field sites used to quantify UHI magnitude, and (2) the measurement variables obtained at those sites. In passing Criterion 2, a study must satisfy both conditions. If an operational definition was not stated in a heat island report, or if the measurement variables or field sites chosen to represent UHI magnitude were not sufficiently explained or illustrated, Criterion 2 failed.

(3) Instrument specifications are explicitly stated in the report, or made implicit through discussion or presentation of data. Instrument specifications include type, mounting and measurement precision.

The WMO is unequivocal in its stance on measurement precision: 'No statement of the results of a measurement is complete unless it includes an estimate of the probable magnitude of the uncertainty', which is normally expressed as the interval of values 'within which the true value of a quantity can be expected to lie' (WMO, 1983). UHI investigators must be explicit in disclosing the measurement precision of their temperature sensors. If measurement precision was stated in a report, as was instrument type, Criterion 3 passed. If instrument type was stated but with no reference to its precision, Criterion 3 failed. Finally, if sufficient detail of instrument mounting (including shielding) was not found in a report, or could not be inferred from its text, tables or figures, Criterion 3 failed.

(4) Site metadata are appropriately detailed in the report. Metadata include a local- or regional-scale map, sketch or photograph of the study area, and one or more quantitative indicators of micro- or local-scale surface exposure, roughness or cover at the field sites used to quantify UHI magnitude.

According to WMO guidelines on climate metadata, all meteorological measurements should include specification of station identity, geographical location, local environment, instrumentation, observing practices, data processing and station history (Aguilar *et al*., 2003). Supplementary WMO guidelines for meteorological measurements in urban areas stress that local environment and historical events are especially important due to the complex and dynamic nature of cities (Oke, 2004). The conditions of Criterion 4 are relaxed from these guidelines, which are too inclusive for a single heat island report. The first condition stipulates that site metadata include a local- or regional-scale illustration (e.g. plan map, site sketch, aerial photograph) of the study area. The illustration must portray major physical and cultural features of the region, such as mountain ranges, valleys, water bodies, transportation routes, built-up areas and other terrain features that are relevant to local and regional surface climate. Also expected of this, or another, illustration are the relative locations of the field sites used to quantify UHI magnitude. The second condition of Criterion 4 stipulates that site metadata include one or more measurable and climatologically relevant indicators of micro- or local-scale surface exposure, roughness or cover of the field sites used to quantify UHI magnitude. Possible indicators include sky view factor, aspect ratio of buildings or trees, fractional coverage of built and natural surfaces, and thermal admittance of built or natural surfaces. If either of these two conditions was not met in a heat island report, Criterion 4 failed. If both conditions were met, Criterion 4 passed.

(5) The micro-scale settings of the field sites used to quantify UHI magnitude are approximately representative, in surface materials, geometry and human activity, of the local-scale surroundings.

The role of scale in Criterion 5 is paramount. UHI investigators are expected to place shelter-height instruments in areas where the local-scale fetch, or 'circle of influence', is relatively homogeneous in surface cover, geometry and human activity. The radius of this circle is difficult to estimate because it changes with building density and atmospheric stability. However, empirical evidence suggests that, as a general rule, the radius is no more than a few hundred metres (Chandler, 1964; Oke, 2004; Runnalls and Oke, 2006). If the micro-scale ($<10^2$ m) setting of a thermal sensor at 1–2 m agl is reasonably uniform, but the local-scale (10^2–10^3 m) surroundings are conspicuously varied or more heterogeneous, then the measured temperatures are not spatially representative, or accurate, beyond the micro-scale area. Investigators who extrapolate temperatures beyond regions of uniformity into wider, more diverse and more complex surroundings are confusing the scales of influence behind their measurements. 'Confusion of scales' is a common flaw in UHI investigation and it amounts to failure of Criterion 5.

In each heat island report, Criterion 5 was judged not on rigorous statistical measures but on qualitative evidence from site maps, photographs, sketches, station names and locations, and descriptions of the study area and its individual field sites. If evidence was sufficient to conclude that investigators used instrument sites approximately representative of the local-scale environment, Criterion 5 passed. If evidence was insufficient to conclude that the sample sites quantifying UHI magnitude were locally representative, the study was graded 'unknown'.

In judging the representativeness of each study's field sites, special attention was given to a controversy known among research reviewers as the 'expectancy effect'. The expectancy effect arises in primary research when investigators induce, through contrived means, a desired or exaggerated response from an experimental test (Hunt, 1997). In empirical UHI studies, the tendency to quantify UHI magnitude with field sites known a priori to exhibit maximum temperature differences, regardless of their representativeness, is a legitimate example of the expectancy effect. Evidence of the expectancy effect in a primary UHI report is adequate warning that field sites may not be representative. Insufficient metadata to allay this warning constitutes failure of Criterion 5.

(6) The number of replicate heat island observations in a report is sufficiently large to meet the stated aims of the study and to yield representative and reliable estimates of UHI magnitude.

Regular and repeated measurement provides control over random variation, and increases the probability of obtaining representative values of a desired effect

at a chosen time and place (Valiela, 2001). Regular measurement also gives reliable basis to inferences. Judgement of Criterion 6 is based on the success with which a study's sample size, or number of repeated heat island observations, is aligned with its aims. Studies boasting large sample sizes were not automatically judged superior to ones with small sample sizes. However, studies with extremely small samples, such as one or a few nights of observation, failed Criterion 6 regardless of their stated aims. If the number of observations in a study could not be found, or could not be deduced from its discussion or presentation of data, Criterion 6 was graded 'unknown'.

(7) The extraneous effects of weather on UHI magnitude are passively controlled. Computations of UHI magnitude use temperatures measured in relatively steady-state weather: no passing fronts, strong advection, or precipitation.

UHI investigators must passively control weather to reduce the risk of confounding 'real' heat islands caused by urban effects with 'fictitious' ones caused by precipitation or air mass advection (Lowry, 1977). Passive control of weather can be gained through preconceived sampling designs or through post hoc data selection. Preconceived sampling avoids frontal or unsettled weather conditions, such as precipitation or strong advection, during data retrieval. Post hoc selection excludes data retrieved during non-steady weather from computations of UHI magnitude, or at least acknowledges weather effects on reported UHI magnitudes. Each of the sample studies was inspected for evidence of non-stationary or unsettled weather in its UHI dataset. If the investigators avoided, removed, or acknowledged the effects of frontal weather in their computations of UHI magnitude, and this effort was explicitly stated, the paper passed Criterion 7. If evidence suggested that frontal weather, especially precipitation and strong advection, had occurred during a measured heat island event, but weather was neither acknowledged as a confounding effect nor excluded from computations of UHI magnitude, the study failed Criterion 7. If neither the observed weather conditions during the heat island events nor any attempts to avoid, remove or acknowledge weather effects were reported, the study was graded 'unknown'.

(8) The extraneous effects of surface relief, elevation and water bodies on UHI magnitude are made sufficiently small through planned sampling design, or made sufficiently known through discussion and recognition of their influences on observed heat island magnitudes.

The effects of surface relief, elevation, and water bodies are difficult to avoid in most UHI studies (Landsberg, 1970; Wanner and Filliger, 1989). Investigators must therefore adopt an appropriate design strategy to counteract unwanted surface influences, otherwise the perceived 'urban heat islands' may not be sufficiently urban-induced to warrant use of this term. Experimental design is critical in eliminating or avoiding the extraneous effects of relief, elevation and water bodies. Placing urban and rural field sites at similar elevation and within relatively uniform local to meso-scale settings is essential for isolating the urban contribution to observed heat islands. Instruments should be sited away from slopes, gullies, cliffs, or ridges, and configured parallel – not perpendicular – to elongated surface features such as valleys and coastlines. These site configurations greatly reduce variable surface effects across a sampled area.

Most urban and rural locations have unwanted surface effects that cannot be avoided, in which case corrective measures can be performed on the data after they have been collected. Two post hoc techniques can improve isolation of the urban effect in complex terrain (Goldreich, 1984). The first technique regresses temperature against height to determine a representative lapse rate for a particular study area. The observed temperatures can then be normalised to a standard level using the measured lapse rates. The second technique regresses temperature against distance inland to determine a representative sea-land profile for a particular study area. Variable sea effects on urban and rural temperatures can then be reduced by normalising the observed temperatures to a standard distance from the shoreline. Both of these post hoc techniques, however, have serious drawbacks – namely, the instability of regression equations – and should be used cautiously, if at all, to correct estimates of UHI magnitude.

Each study was assessed of its success with Criterion 8 on evidence gathered from its discussion and illustration of the study area and on the individual field sites used to quantify UHI magnitude. If, through planned sampling, UHI investigators were unable to avoid the disturbing surface features of a particular study area, they should instead account for the surface factor in other ways. At minimum, they should qualify their estimates of UHI magnitude by appropriately recognising unwanted surface effects on measured heat island magnitudes. Recognition of these effects may include one or more of the post hoc regression techniques previously discussed. Post hoc correction by itself, however, does not constitute a passing grade – it must be part of a broader treatment of the surface factor that qualifies the purported 'urban' heat island estimates as over- or under-estimates by way of unavoidable land-surface features.

Given the difficulty and uncertainty of establishing control over the effects of surface relief, elevation and water bodies on UHI magnitude, qualitative treatment alone of the topo-climatic effect constitutes a passing grade for Criterion 8. Similarly, if, by planned sampling design, investigators sufficiently reduced or eliminated surface relief, elevation and water body effects from measured UHI magnitudes, passing grades were earned. Investigations that disregarded extraneous surface effects altogether from their study areas, and thus failed to discriminate a reasonably accurate urban factor, met Criterion 8 unsuccessfully. If a report did not describe

or depict the surface features of a study area in sufficient detail, or did not disclose the locations of the field sites used to quantify UHI magnitude, Criterion 8 was graded 'unknown'.

(9) Temperatures used to quantify UHI magnitude are measured synchronously. Inhomogeneities resulting from non-synchronous measurement are acknowledged as such and adjusted to a common base time.

Criterion 9 highlights the importance of time control during UHI measurement. If the temperatures used to quantify UHI magnitude are not synchronous, or adjusted so as to be synchronous, *urban-induced* heat islands may be confounded with *time-induced* heat islands. If regional temperature change during mobile data-collection was said or shown to be significant by the investigators, and temperature-time adjustments were carried out, Criterion 9 passed. If temperature-time adjustments were judged to be necessary, but were not acknowledged in the investigation, Criterion 9 was graded 'unknown'. Investigations that used temperature minima to quantify UHI magnitude were likewise expected to apply temperature-time corrections to their data. Temperature minima yield unreliable estimates of UHI magnitude because they are not normally synchronised across a spatial network of instruments, especially over complex urban and rural topography or in non-steady weather (Oke and Maxwell, 1975; Szymanowski, 2005). Investigations that failed to acknowledge or execute temperature-time corrections, which were judged to be necessary, did not pass Criterion 9.

2.3.2. *Grading scheme*

A points-based grading scheme was designed to quantify methodological quality in the heat island literature sample. Each sample report was graded and ranked by a conventional 'vote count' procedure in which points were awarded for passing a criterion and no points for failing a criterion (Glass, 1976). A study earned a maximum of 18 points for passing all nine scientific criteria. The number of points assigned to each criterion is based on its weight in generating reliable and reasoned estimates of UHI magnitude (Table I). Criteria 1, 2 and 5 are deemed 'critical' to a reliable UHI estimate and weigh heavily in the grading scheme. Criteria 3, 4 and 6 are deemed unnecessary – but still 'desirable' – and consequently weigh less. Criteria 7–9 are deemed 'somewhat essential' and carry intermediate weight. The grading scheme also allowed partial points for specific criteria if their antecedent conditions were successfully met.

Table I. Points-based grading scheme for assessing methodological quality in the heat island literature sample.

Criterion	Weight class	Total points allotted	Points allotted by grade			Partial points
			Fail	Unknown	Pass	
1. Conceptual model	Critical	3	0	0	3	No
2. Operational definitions	Critical	3	0	–	3	No
3. Instrument specifications	Desirable	1	0	–	1	One-quarter point each for mounting and shielding; one-half point for precision
4. Site metadata	Desirable	1	0	–	1	One-half point each for site map and quantitative indicator
5. Site representativeness	Critical	3	0	0	3	No
6. Number of replicates	Desirable	1	0	0	1	No
7. Weather control	Somewhat essential	2	0	0	1 or 2	1 point for post hoc treatment; 2 points for planned sampling
8. Surface control	Somewhat essential	2	0	0	1 or 2	1 point for post hoc treatment; 2 points for planned sampling
9. Synchronicity	Somewhat essential	2	0	0	1 or 2	1 point for near-synchronous measurement or temperature-time correction; 2 points for synchronous measurement
Total	...	18	0	0	15–18	–

The sample studies were then sorted into three tiers based on their overall success with the nine scientific criteria. *Top tier* studies and their estimates of UHI magnitude earned 11–18 points and are of the highest methodological quality in the literature sample. These studies follow the scientific method to the extent that the conceptual model and operational test are aligned, operational definitions are clearly stated, field sites are approximately representative of the local environment, and extraneous influences on measurement are carefully controlled. Only those studies near the top of the points range give full account of instrument specifications and site metadata, and gather a sufficiently large sample to control random variation. *Middle tier* estimates of UHI magnitude earned 7–15 points toward their quality scores and are acceptable only on the condition that certain weaknesses or uncertainties in method are acknowledged. *Bottom tier* estimates of UHI magnitude earned only 1–12 points and are deemed unacceptable. These studies are crudely designed and yield methodologically unsound or unreliable UHI estimates regardless of their success with the scientific criteria. With insufficient control of confounding effects like weather, relief and time, the reported UHI magnitudes are induced as much through non-urban effects as through urban effects. Bottom tier studies are consequently at high risk of attributing false cause to observed heat island magnitudes.

Each study moved through a criteria-based scheme to determine its appropriate tier placement (Figure 2). Standards for tier placement are most demanding in the top tier and least demanding in the bottom. Studies were assigned to the tiers based on their success with only the 'critical' and 'somewhat essential' criteria. The 'desirable' criteria had no bearing on tier placement and were used only for determining points within tiers. Studies with similar point totals but different tier placements are alike only in the quantity of criteria passed, not in the combination of criteria passed.

During the grading process, each study was tagged with a missing data index (MDI) measuring its completeness and efficiency of reporting (Pigott, 1994). MDIs were determined by tallying the number of points lost to 'unknown' grades, which was then converted to a percentage of the total number of 'unknown' points available (13). MDI values were normalised from 0 to 1. Values approaching unity indicate a detrimental lack of information in a report, and raise the possibility of unconventional or unrepresentative instrument siting and/or lack of experimental control. Values approaching zero indicate full and competent reporting. One might argue that studies with excessive reporting gaps (i.e. many 'unknown' grades) should be removed from a systematic review because they cannot be rated fairly against those with more complete reporting. Given that reporting itself is a measure of research competence, the argument to remove studies that are weak in communication is immaterial to this review's aims and desired output.

Rankings: In the final stage of evaluation, the quality scores of each primary study were converted to rank equivalents. Rankings were determined first by tier placements and second by quality scores. Accordingly, studies in the top tier were ranked above those in the middle and bottom tiers, and studies with high scores above those with low scores. If two or more studies had identical tier placements and scores, the study earning more points from the three 'critical' criteria was ranked higher. If the studies earned an equal number of points from the 'critical' criteria, the study earning more points from the 'somewhat essential' criteria was ranked higher. Studies

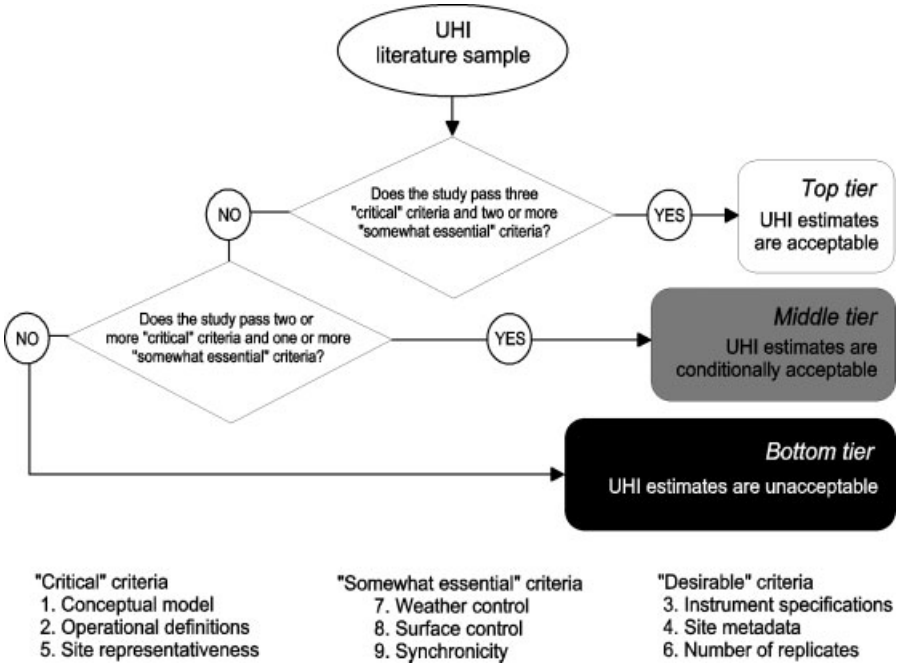

Figure 2. Criteria-based scheme for determining tier placements in the heat island literature sample.

still equal in point earnings from the 'somewhat essential' criteria were assigned shared ranks.

3. Results

3.1. Describing the literature sample

More than 500 candidate papers and online articles and abstracts were screened for inclusion in the systematic review. Of this total, 177 papers were declared eligible for assessment. A total of 88 stationary and 102 mobile subsamples were extracted from these eligible studies, giving an aggregate sample size of 190 heat island studies. The number of eligible papers and the study sample size are different because papers classified by method of data collection as both 'mobile' and 'stationary' were graded twice.

The heat island observations reported in the literature sample are distributed across 11 continental realms and 221 cities and towns (Figure 3). In more than half of the 177 sample papers, the observations originate from European and North American cities, and in one-quarter of the papers they originate from East and South Asian cities. The remaining seven geographic realms are each represented by ten or fewer papers. Continental realms having a larger percentage of the sample's total urban population are not necessarily represented by greater frequencies of heat island papers (Figure 4). Europe and North America, for example, are overrepresented in the sample, whereas North Africa, Southwest Asia, South America, and Middle America are all underrepresented. Geographic breakdown by political region puts the United States in the frequency modal class, with 29 papers, followed by the United Kingdom (20), Japan (17), Canada (15) and India (12). Seven foreign languages are represented in the literature sample: English is the modal class, with 152 papers, followed by Japanese (8), German (5), Chinese (5), Spanish (3), French (2), Russian (1) and Korean (1). Frequency distribution by year of print or publication is positively skewed across the 58-year sample period (Figure 5). The first and last decades are represented by the lowest and highest frequencies, respectively, in the study sample. Frequency values range from 5 studies between 1950 and 1959, to 49 studies between 2000 and 2007.

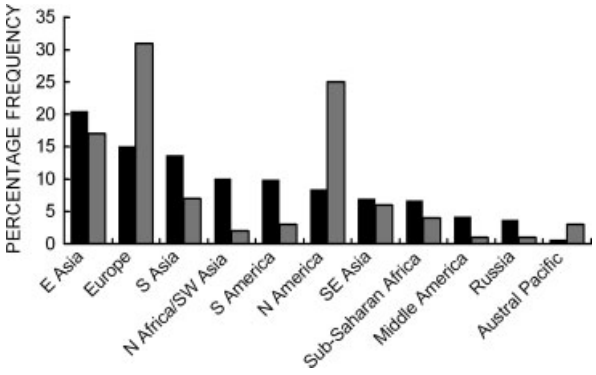

Figure 4. Percentage frequency distribution of the literature sample ($n = 177$) by geographic realm, urban population (black bars), and number of heat island studies (grey bars).

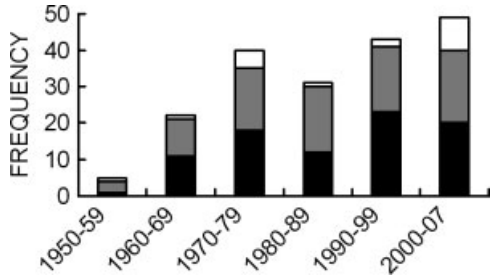

Figure 5. Frequency distribution of the heat island literature sample ($N = 190$) by decade and tier placement. Black bars = bottom tier; grey bars = middle tier; open bars = top tier.

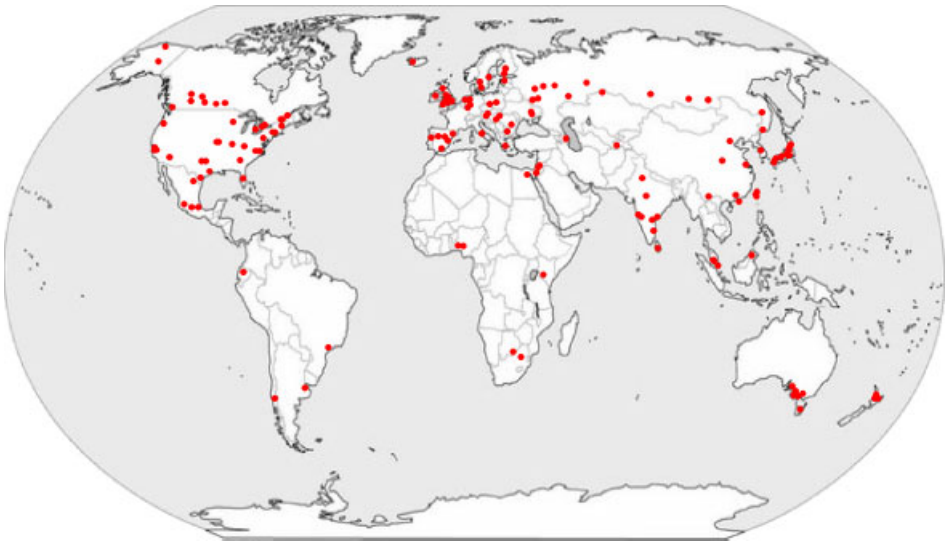

Figure 3. Geographic distribution of heat island observations in the literature sample. This figure is available in colour online at wileyonlinelibrary.com/journal/joc

The literature sample was retrieved from a variety of sources. The highest frequency of papers, at 68% of the total sample, were retrieved from peer-reviewed scholarly journals. Non-refereed academic journals follow at 12%, and non-refereed professional/trade journals at 9%. The remainder of the sample is comprised of government reports, institute papers, technical notes, book chapters and magazines articles. By method of data collection, the literature sample comprises 89 'mobile' and 75 'stationary' studies. The mobile studies primarily used automobiles to transport their temperature sensors across an urban-rural area, although trains, motor-scooters and bicycles were also used. Stationary studies used *in situ* or purpose-built networks of urban and rural temperature sensors. Thirteen papers in the literature sample used both mobile and stationary surveys to quantify UHI magnitude.

3.2. Analysing the primary literature

A full summary of the pass and fail ratios for the nine scientific criteria is provided in Table II and Figure 6. Criteria 1 (Conceptual model) and 2 (Operational definitions) have the highest aggregate pass ratios, at 75 and 78%, respectively, of all nine criteria. Twenty-three percent of the 190 sample studies were graded 'unknown' for their conceptualisation of the UHI effect. Twenty-two percent of the sample studies failed Criterion 2, meaning that nearly one-quarter of all studies in the literature sample provide no definition or explanation of UHI 'magnitude' or 'intensity', nor any evidence on which to base a reasonable inference of that definition.

Despite an aggregate passing ratio of 75% for Criterion 1, the discrepancy between ratios for the mobile and stationary sub-samples is large. Almost all mobile studies (97%) passed Criterion 1 compared to only half (49%) of stationary studies. The success rate for mobile studies

Table II. Pass and fail ratios by scientific criterion and method of data collection.

Criterion	*n*	No. of 'passing' grades	No. of 'failing' grades	No. of 'unknown' grades	'Passing' ratio (%)	'Failing' ratio (%)	'Unknown' ratio (%)
1. Conceptual model							
Mobile	102	99	1	2	97	1	2
Stationary	88	43	3	42	49	3	48
Aggregate	190	142	4	44	75	2	23
2. Operational definitions							
Mobile	102	72	30	–	71	29	–
Stationary	88	77	11	–	88	12	–
Aggregate	190	149	41	–	78	22	–
3. Instrument specifications							
Mobile	102	33	69	–	32	68	–
Stationary	88	10	78	–	11	89	–
Aggregate	190	43	147	–	23	77	–
4. Site metadata							
Mobile	102	12	90	–	12	88	–
Stationary	88	9	79	–	10	90	–
Aggregate	190	21	169	–	11	89	–
5. Site representativeness							
Mobile	102	9	13	80	9	13	78
Stationary	88	15	27	46	17	31	52
Aggregate	190	24	40	126	13	21	66
6. Number of replicates							
Mobile	102	40	60	2	39	59	2
Stationary	88	76	10	2	87	11	2
Aggregate	190	116	70	4	61	37	2
7. Weather control							
Mobile	102	72	2	28	71	2	27
Stationary	88	32	53	3	36	60	4
Aggregate	190	105	54	31	55	29	16
8. Surface control							
Mobile	102	57	19	26	56	19	25
Stationary	88	46	18	24	52	21	27
Aggregate	190	103	37	50	54	20	26
9. Synchronicity							
Mobile	102	82	7	13	80	7	13
Stationary	88	54	33	1	61	38	1
Aggregate	190	136	40	14	72	21	7

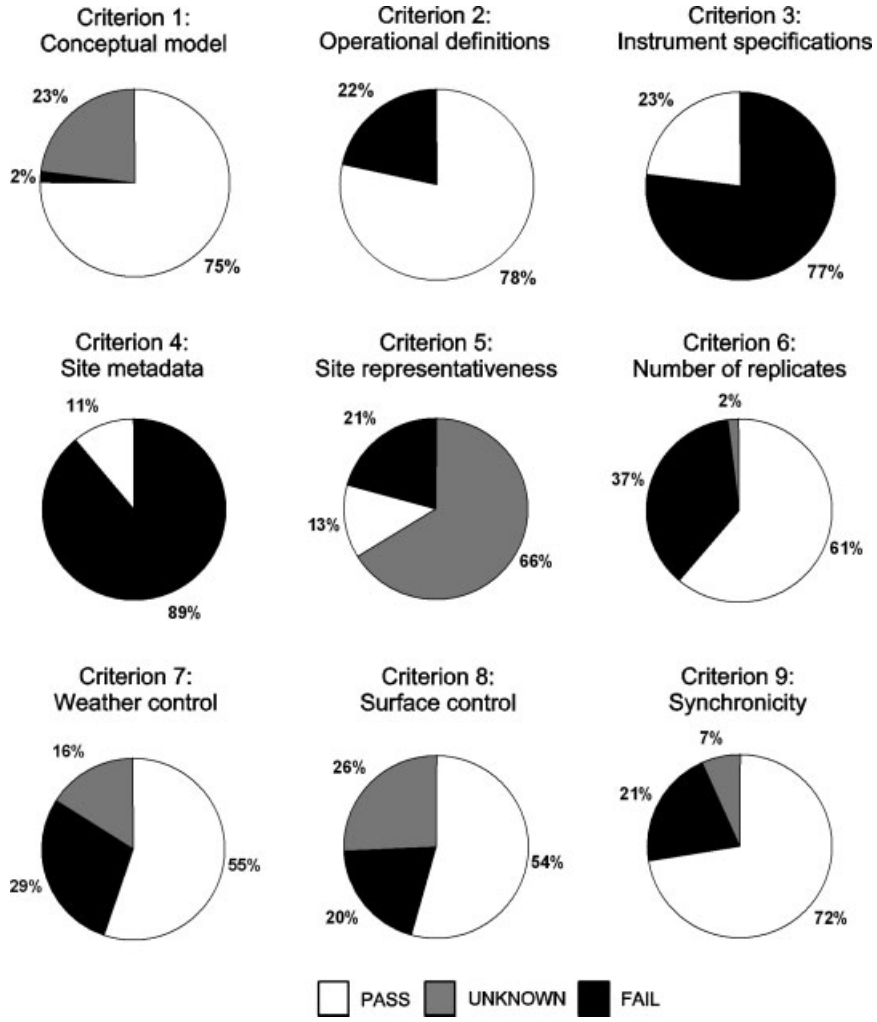

Figure 6. Frequency distribution of the heat island literature sample ($N = 190$) by scientific criterion and aggregate pass/fail ratios.

is high. Simple description of their modes of transport (e.g. automobiles or motor-scooters) makes implicit the fact that air temperature measurements were made in the canopy-layer. Clear statements of instrument height are therefore less critical to the judgment of Criterion 1. In contrast, the use of stationary surveys involving fixed weather boxes or towers requires more explicit description from investigators as to instrument height relative to the canopy-layer. For this reason, the frequency of 'unknown' grades in the literature sample is disproportionately higher in the stationary sub-sample, at 48%, than in the mobile sub-sample, at 2%. The stationary passing ratio for Criterion 2 is slightly higher than the mobile ratio, suggesting a greater tendency for these investigations to make their operational definitions of UHI magnitude known. With fewer measurement sites, on average, stationary studies tend to have simpler and more communicable definitions of UHI magnitude than mobile surveys.

Aggregate pass and fail rates for Criteria 3 (Instrument specifications) and 4 (Site metadata) are inversely proportional to those of 1 and 2. At 77 and 89%, respectively, Criteria 3 and 4 have the highest failure rates of all nine criteria. Only 43 of a total 190 studies, or 23% of the sample, provide full details of their instruments. Of the 147 studies that fail to give full details, 97% gives no indication of precision, 39% no indication of type and 40% no description of mounting or shielding. Of the 169 studies that failed Criterion 4, only 8% failed on account of incomplete or incompetent cartographic or photographic representation of the study area. Ten percent of the studies failing Criterion 4 were unsuccessful because the major physical and cultural features influencing local surface climate in the study area were not depicted in regional maps or illustrations, whereas 17% give no depiction whatsoever of the field site locations defining UHI magnitude. Accounting for a much larger fraction of the failing grades in Criterion 4 is the deficiency of quantitative descriptors of micro- or local-scale site character. Of the 169 studies failing Criterion 4, 168 provide no quantitative description of the field sites defining UHI magnitude. Most of these studies instead use qualitative expressions like 'green fields' or 'city centre' to describe their sites and local settings. Thirty-three percent of the literature sample gives neither qualitative nor quantitative descriptions of their field sites and settings. These studies use only 'urban' and 'rural', or other equally vague terms, to describe their sites. In other

studies, quantitative descriptors are provided for one or several of the sites defining UHI magnitude, but not for all sites. The most frequently cited quantitative descriptor of site character in the primary literature is fractional coverage of built and natural surfaces, followed by the height of roughness elements (e.g. buildings, trees), and, finally, sky view factor. These descriptors are each cited in less than 10% of the 190 sample studies.

The high failure rate for Criterion 4 (Site metadata) has negatively influenced the outcome of Criterion 5 (Site representativeness). Sixty-six percent of the literature sample has field sites of unknown representativeness, largely because these studies are lacking site metadata. One-fifth of the total sample was judged on sufficient evidence to have unrepresentative field sites quantifying UHI magnitude, while only 13% of the literature sample provides sufficient description of their sites to earn passing grades for Criterion 5. Barring any convincing evidence for or against site representativeness, studies using fixed-interval or grid sampling techniques for site selection were graded 'unknown'. These techniques give no consideration to the climatological character of the surfaces at the chosen sites. Six of the forty studies failing Criterion 5 openly confess that their field sites are not representative of the local surroundings, and give evidence to support these claims. The ratios of 'unknown' grades for Criterion 5 are significantly different between the stationary and mobile sub-samples. Half of the 88 stationary studies were judged 'unknown' for site representativeness, whereas over three-quarters of the 102 mobile studies were judged 'unknown'. This discrepancy is in part a consequence of the much lower spatial resolution of temperature sampling associated with stationary data collection. The likelihood of a stationary study describing the character of its sites in sufficient detail to pass Criterion 5 is therefore greater than that of a mobile study.

In contrast to Criteria 1–5, the remaining criteria were met with moderate success. Aggregate pass rates for Criteria 6 (Number of replicates), 7 (Weather control) and 8 (Surface control) range from 54 to 61%. Criterion 6 has a passing ratio of 61%, meaning that the majority of sample studies gathered a sufficiently large number of observations on which to base reliable inferences of UHI magnitude. Thirty-seven percent of the sample was judged unsuccessful in carrying out observations of sufficient duration or frequency, or in meeting the stated aims of their investigation. Fifty-nine percent of these failing studies involved mobile surveys, and only eleven percent stationary surveys. Mobile surveys are comparatively labour- and resource-intensive and are therefore disadvantaged by data of poor temporal resolution. Stationary surveys, in contrast, are operationally simple and thus favoured for replicate, frequent and long-term observations of UHI magnitude. Studies graded 'unknown' comprise just 2% of the total sample.

Criteria 7 (Weather control) and 8 (Surface control) met similar passing ratios of about 55%. The remaining 45% of the sample studies failed to sufficiently control – through planned sampling design or post hoc data correction/selection – the disturbing effects on UHI magnitude, or to communicate the extent to which control was taken. In either case, nearly half of the reported UHI magnitudes in the literature sample were judged to be confounded beyond acceptability by non-urban effects on temperature. Only 2% of mobile studies failed Criterion 7, compared to 59% of stationary studies. The significantly lower failure rate in the mobile sub-sample is explained by the freedom that investigators have in controlling the time of a mobile survey. If weather effects potentially distort the measured heat island signal, investigators can abandon or delay data collection until more desirable conditions develop. Stationary surveys, however, require investigators to manually remove the distorting effects of weather from their data sets because the heat island signal is recorded continually through all weather conditions.

More than half of the 190 sample studies successfully met the conditions of Criterion 8 (Surface control). Success with Criterion 8 is predicated not only on the investigators' calculated attempts to reduce surface effects on UHI magnitude, but also on the surface complexity of the area in which these attempts are carried out. Of the 103 studies that passed Criterion 8, 54% succeeded on account of planned sampling design and 46% on post hoc data correction. The 37 studies that failed Criterion 8 were judged inadequate in their attempts to recognise and separate *surface* and *urban* influences in their estimates of UHI magnitude. The percentage of 'unknown' grades in Criterion 8 is relatively high, at 26%.

Criterion 9 (Synchronicity) has pass and fail ratios of 72 and 21%, respectively. All studies that failed Criterion 9 derived UHI estimates from non-synchronous termperature measurements. The remaining 7% of the sample was graded 'unknown' for incomplete reporting of temperature/time data, or of attempted corrections to those data. In these investigations, the duration of the mobile surveys is not reported and thus the extent to which temperature-time corrections are needed is not known. The failing rates for the stationary and mobile sub-samples in Criterion 9 are significantly different, at 38 and 7%, respectively. Stationary studies that failed Criterion 9 use temperature minima to quantify UHI magnitude, wherease mobile studies that failed Criterion 9 use temperature data that were left unadjusted for temporal inhomogeneities despite survey times lasting several hours.

3.3. Grading the primary literature

3.3.1. Tier placements and quality scores

Frequency distribution by tier placement favours the bottom and middle tiers of the grading scheme (Figure 7). Forty-five percent of the sample studies were placed into each of the middle and bottom tiers, with the remaining 10% placed in the top tier. The distribution of studies

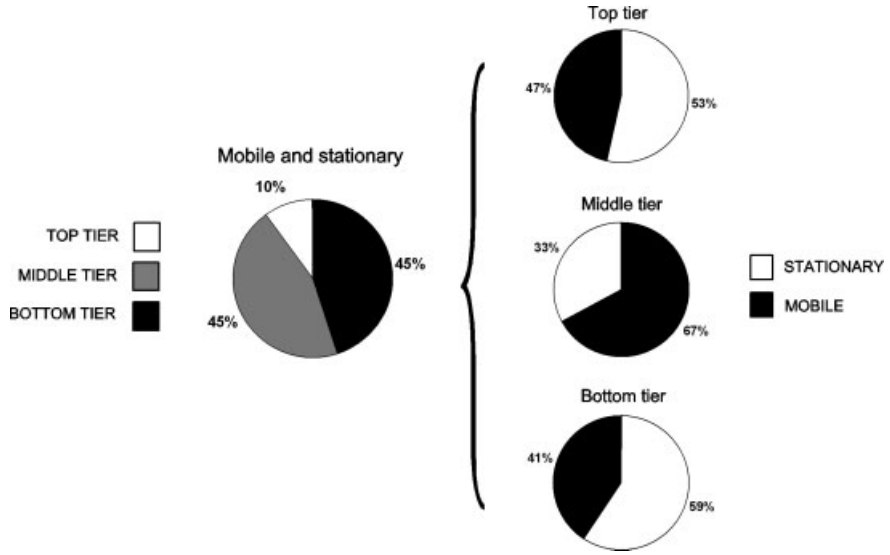

Figure 7. Frequency distribution of the heat island literature sample ($N = 190$) by tier placement and method of data collection.

across the three tiers changes slightly between the mobile and stationary sub-samples. The majority of studies in the bottom and top tiers are stationary, whereas in the middle tier the majority are mobile.

Quality scores for the study sample range from 1 in the bottom tier to 18 in the top tier (Figure 8). Only 2 of 190 studies earned maximum scores of 18. The mean quality score for the entire sample is 9.3, and the modal class is the 10 to <11 point range, with 25 of 190 studies. The distribution of scores around the mean is symmetrical, with fewer studies in higher and lower point ranges. The mean scores for the bottom, middle and top tiers are 6.3, 10.8, and 15.5, respectively (Table III). Distribution of tier placements by decade reveals that almost half of the top tier studies were printed or published between 2000 and 2007 (Figure 5). More than one-quarter of the bottom tier studies were printed or published between 1990 and 1999.

MDI values in the UHI literature sample range from 0 to .85 (Figure 9). The frequency distribution is skewed slightly to the left, indicating that high MDI values are

Table III. Mean quality scores and missing data index (MDI) values by tier placement and method of data collection.

Tier	*n*	Mean quality score[a]	Mean MDI value
Top			
Mobile	9	14.6	.09
Stationary	10	16.4	.02
Aggregate	19	15.5	.05
Middle			
Mobile	58	10.6	.28
Stationary	28	11.2	.19
Aggregate	86	10.8	.25
Bottom			
Mobile	35	6.9	.36
Stationary	50	5.9	.39
Aggregate	85	6.3	.38
All			
Mobile	102	9.6	.29
Stationary	88	8.8	.28
Aggregate	190	9.3	.29

[a] Out of 18 points.

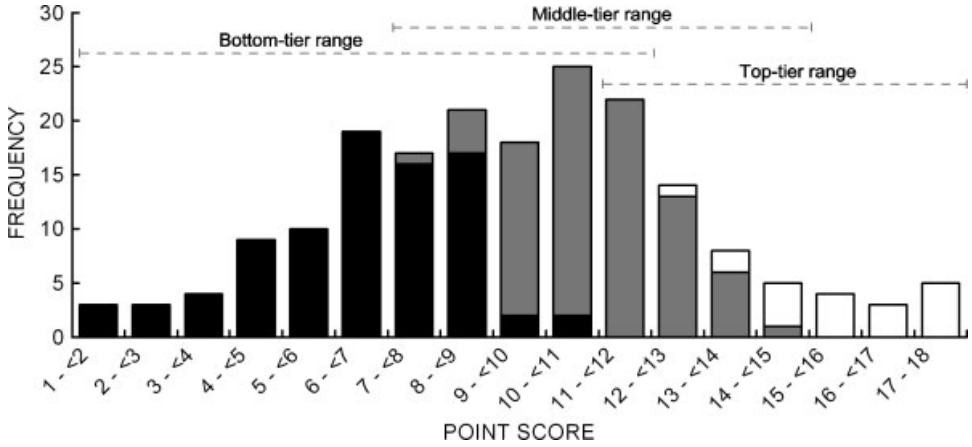

Figure 8. Frequency distribution of the heat island literature sample ($N = 190$) by point-based quality scores and tier placement. Black bars = bottom tier; grey bars = middle tier; open bars = top tier. Mean quality score = 9.3.

less frequent than low values. The frequency modal class is the .2 to <.3 range of values. Sixteen percent of the sample has MDI values of 0. Communication in these studies, which belong entirely to the top tier, is complete and grading was unimpaired by missing information. Twelve percent of the sample has MDI values of .05 or greater. In these studies, which belong mainly to the bottom tier, more than half of the information needed to pass Criteria 1 and 5–9 was missing. No studies in the sample were assigned MDI values of 1. The mean MDI value for the literature sample is .29, and for the top, middle and bottom tiers, mean values are .05, .25, and .38, respectively (Table III).

3.3.2. *Rankings*

Rankings, MDI values, quality scores and titles of the top tier studies in the sample are listed in Appendix A (Table AI). The correlation between rank numbers and MDI values is negative ($r = -.64$), suggesting that high rank/tier placements are associated with efficient reporting, and low rank/tier placements with incomplete or incompetent reporting (Figures 9 and 10). The spread of values along the vertical axis of Figure 10 shows that studies with MDI ratings of 0 range in rank placements from 1 to 142. As MDI ratings increase, the range in rank placements greatly diminishes, meaning that incomplete reporting has a stronger influence on rank placement than complete reporting. The explanation for this pattern is that complete reporting does not guarantee a study's success with the scientific criteria, whereas incomplete reporting necessarily guarantees a study's poor performance with the criteria.

4. Discussion

Before reflecting on the results of the systematic review, I offer several caveats to their interpretation. These caveats are meant only to improve the readers' understanding of the grades, scores and rankings and not to excuse or justify them. I then comment on the overall quality of the empirical UHI literature, as measured by the systematic review and its statistical output, and identify areas for generalisation regarding its methodological strengths and weaknesses. I close the paper with recommendations for improving the literature and promoting a critical perspective on UHI observation and reporting.

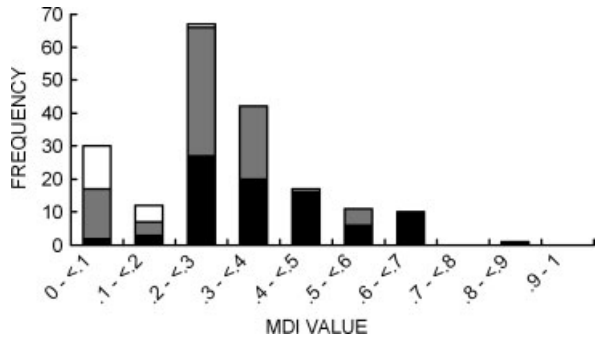

Figure 9. Frequency distribution of the heat island literature sample ($N = 190$) by missing data index (MDI) values and tier placement. Black bars = bottom tier; grey bars = middle tier; open bars = top tier. Mean MDI value = .29.

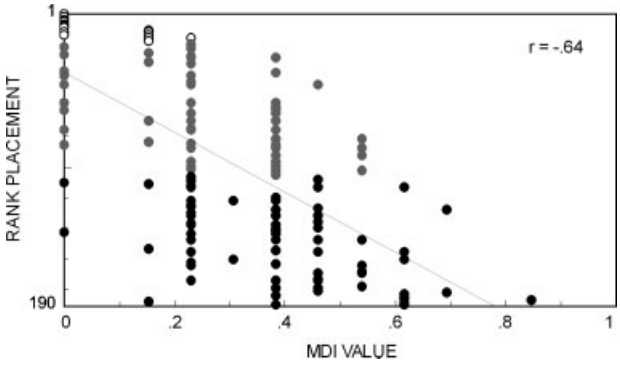

Figure 10. Distribution of the heat island literature sample ($N = 190$) by missing data index (MDI) values, rank placement, and tier placement. Black circles = bottom tier; grey circles = middle tier; open circles = top tier.

4.1. Caveats

This review was conducted with as much objectivity and uniformity as possible. Its output, however, is ultimately based on the knowledge, skilled judgement and critical imagination of a single reviewer. The review is therefore inherently subjective and in no way reflects the views of other colleagues, collaborators or contributors. Many layers of subjectivity exist in the review process, from the initial selection of literature to the final grading of its content. Hidden among these layers are the collective experiences and preconceived beliefs of all previous authors represented in the literature sample. In reducing these biases to the extent possible, each paper was read with a mindset clear of expectations for particular authors, institutions or regions, and with equal respect and fairness toward the researchers and their reported findings.

The results of the systematic review should be interpreted with consideration for the challenges that all UHI investigators face. Foremost among these challenges is the complexity of natural settings in which the observations are conducted. Thus the grades, scores, tiers and rankings assigned to each sample study reflect not only its methodological quality, but the success of its investigators in designing a measurement program that is best suited to the natural setting of the study area and that is financially and technologically feasible. For this reason, scores and rankings may vary among studies that appear similar in methodology but different in natural setting or in technical or financial backing. Scores and rankings should not be construed as judgements on the personal or professional competence of the investigators, or on aspects of a paper not related to empirical estimation of UHI magnitude.

4.2. Extracting generalisations from the systematic review

Overall, the quality of the UHI literature and its empirical content is low at best. The mean quality score for the study sample is just 52%, and nearly half of the evaluated studies provide estimates of UHI magnitude that are unacceptable in terms that environmental science can reasonably expect. Many of these studies report observations that are too casual to identify and isolate a proper UHI effect. Furthermore, the study sample is missing, on average, nearly one-third of the information needed to fully assess the methodological quality of its UHI estimates, and thus a significant portion of the literature has empirical content of unknown or indeterminate standing. These results expose a literature that is lacking rigour in most aspects of experimental design and communication, and that will not gain the trust of a discerning reader. The larger implication in these findings is that less is known about the UHI magnitudes of cities worldwide than might be anticipated from a literature of such historical and geographical breadth.

Although this outlook may be a discouraging one, consolation lies in the finding that a small portion of the literature sample is outstanding in its approach to UHI estimation. These studies report UHI observations and estimates that are focused and systematic, that face few threats to their validity, and that are fully acceptable within the constraints of environmental science. As such, they provide high standards for designing and judging future heat island investigations, and should be valued for this purpose. These studies also provide reliable input data for use in generalised boundary-layer models, in algorithms for predicting heat island magnitude, and in data-correction schemes for removing urban bias from regional and global climate assessments. Further consolation lies in evidence of a 'learning effect' in the literature sample. Nearly half of the top tier studies in the sample were conducted in the past decade alone, while all remaining top tier studies were conducted in decades prior. This implies that the quality of UHI estimates reported in the literature is recovering through time as understanding of heat islands advances and observational techniques improve.

In assessing the methodological strengths and weaknesses of the literature, I identify three areas for generalisation. These generalisations are based on a study sample that is both large and homogeneous, and thus I am confident that my remarks are valid across a wider population of ground-based canopy-layer heat island observations. My remarks are less valid, however, outside of this population and I caution readers against extending the generalisations to remotely sensed, boundary-layer, and non-urban heat island studies.

The first area for generalisation is *operationalisation of concepts*. The literature is reasonably successful in this regard, as most studies demonstrate good conceptual understanding of the heat island effect and establish appropriate definitions to test these concepts. Especially encouraging is the placement of instruments at proper heights for measuring canopy-layer UHI magnitude. Still, concern lingers over a minority of studies that fail to specify instrument height or define UHI magnitude in operational terms.

The second area for generalisation is *controlled measurement*. The literature is generally poor in this regard. Approximately half of all heat island studies fail to sufficiently control their measurements for the confounding effects of weather, relief or time. With no ability to discriminate urban from non-urban effects on temperature, these studies easily confuse meaningful results with chance results. Control over weather is especially problematic in stationary surveys, as few investigators filter or correct their data for its disturbing effects. Time control is well executed in most mobile surveys, but is problematic in stationary surveys.

The third area for generalisation is *openness of method*. The literature is highly inadequate in this area, with three-quarters of the sample failing to communicate, in most basic terms, the precision of instruments used to measure UHI magnitude and the physical nature of the surfaces surrounding those instruments at the time of measurement. Incompetent reporting of site metadata in turn makes meaningful communication of site representativeness difficult or impossible. Mobile studies are especially guilty because the number of field sites used to quantify UHI magnitude is often too large to fully account for their surface character. Openness of method is a lesser concern in other areas of communication, although definition of UHI magnitude and control of measurements is frequently not reported.

4.3. Recommendations and closing remarks

In closing, I offer the following recommendations for improving methodological quality in heat island studies and their estimates of UHI magnitude. These recommendations are intended to promote better communication and understanding among all researchers of the heat island effect, and to provide a critical framework for assessing future heat island reports.

1. *Reduce the spatial and temporal resolution of your data.* For the purpose of quantifying UHI magnitude, fewer field sites in representative locations is preferable to more sites in unrepresentative locations. Likewise, a smaller dataset of controlled measurements is preferable to a larger dataset of uncontrolled measurements. A simple comparison of two representative sites will provide a reasonably good measure of UHI magnitude, provided that the measurements sufficiently regulate the effects of weather, relief, time and random variation. Fewer sites and replicate observations in turn simplify control and communication of procedures. Stationary instruments that are automated and synchronised are immediately advantaged over mobile surveys.
2. *Follow standardised guidelines for site reporting.* Guidelines in Aguilar *et al.* (2003) and Oke (2004) include descriptive templates for reporting the micro-,

local- and meso-scale settings of temperature measurements in urban and rural environments. The information contained in these templates is essential to any heat island paper and to proper interpretation and comparison of its reported UHI magnitudes. The proposed site classification system of Stewart and Oke (2009a, 2009b) is also a useful tool for site reporting because the communication of physical site properties is explicit in its portrayal of built and natural landscapes. Most of the metadata needed for site reporting can be obtained first-hand at the field sites themselves, or second-hand from meteorological offices, local observers/experts, libraries (e.g. historical photographs, maps), or online portals for digital imagery and mapping (e.g. Google Earth/Maps).

3. *Disclose the limits of your data.* Observational data in environmental science are limited in their certainty and reliability. Like all climate observations, UHI measurements are limited by the complexities of the surface–atmosphere system and by the technical capacity of our instruments to sample that system. Public statements claiming exact and absolute values of UHI magnitude are unjustified because the phenomenon being measured is inherently complex and difficult to access. Honest reporting of limitations and errors in observation is the best practice for sharing and advancing knowledge of UHIs. Public statements should instead claim 'reasonable estimates' of UHI magnitude, and couch these estimates in round figures, within margins of instrumental error, and with a tone of caution.
4. *Use terminology with discretion.* The term "urban heat island" is used irresponsibly in the literature to describe all observed city-country temperature differences regardless of the causes behind those differences. If the temperature differences in a particular city are caused primarily by weather or topographic interferences, then the perceived heat island should not be described as an urban-induced one. "Urban heat island" should instead be reserved for observations that have been sufficiently controlled for non-urban influences. Discretionary use of this term will further promote control of measurements.
5. *Never accept UHI magnitudes at face value.* Behind every reported estimate of UHI magnitude is an extenuating set of circumstances. These circumstances are both experimental (e.g. definition, instrumentation and measurement) and environmental (e.g. weather, climate and topography). No estimate of UHI magnitude is of any value to the public unless its extenuating circumstances are fully disclosed. Public comparison of UHI magnitudes in the literature is risky because these circumstances are often not reported or properly understood. Especially risky is the unqualified comparison of UHI magnitudes based on population or land use.

These five recommendations call on the critical minds of research reviewers and heat island investigators to scrutinise the literature, weigh its results and ultimately question its validity. Awareness, critique and revision of method are important stages in this process, as is demand for reduced but more responsible reporting of primary research. If climate modellers, weather forecasters, city planners, urban engineers and building architects are to be convinced of the serious environmental and social implications behind the UHI effect, heat island researchers must first produce results that can be trusted.

Acknowledgements

I thank the volunteer translators who worked long hours transcribing foreign-language materials used in this research. I also thank Professor Tim Oke (University of British Columbia) for loaning historical documents from his urban climate archive, and for offering helpful comments on this paper and its original manuscript. Professor Michael Church (University of British Columbia) also offered helpful suggestions for improving the methodology and presentation of this paper. This research is funded by a Discovery Grant to Tim Oke and a Doctoral Fellowship to I. D. Stewart from the Natural Science and Engineering Research Council of Canada.

References

Aguilar E, Auer I, Brunet M, Peterson TC, Wieringa J. 2003. *Guidance on Metadata and Homogenization*. World Meteorological Organization: Geneva. WMO Technical Document No. 1186.

Arnfield J. 2003. Two decades of urban climate research: A review of turbulence, exchanges of energy and water, and the urban heat island. *International Journal of Climatology* **23**: 1–26.

Bohm R, Gabl K. 1978. The urban heat island in dependence of different meteorological parameters. [In German]. *Archives for Meteorology, Geophysics, and Bioclimatology B* **26**: 219–37.

Brooks CEP. 1952. Selective annotated bibliography on urban climates. *Meteorological Abstracts and Bibliography* **3**: 734–773.

Chandler TJ. 1962. Temperature and humidity traverses across London. *Weather* **17**: 235–241.

Chandler TJ. 1964. City growth and urban climates. *Weather* **19**: 170–171.

Chandler TJ. 1970. Urban climatology – Inventory and prospect. In *Urban Climates – Proceedings of the Symposium on Urban Climates and Building Climatology, October 1968, Brussels*. WMO Technical Note No. 108. World Meteorological Organization: Geneva.

Cooper H, Hedges LV. 1994. Research synthesis as a scientific enterprise. In *The Handbook of Research Synthesis*, Cooper H, Hedges LV (eds). Russell Sage Foundation: New York.

Glass GV. 1976. Primary, secondary, and meta-analysis of research. *Educational Researcher* **5**: 3–8.

Goldreich Y. 1984. Urban topoclimatology. *Progress in Physical Geography* **8**: 336–364.

Hempel CG. 1966. *Philosophy of Natural Science*. Prentice-Hall: Englewood Cliffs, NJ.

Howard L. 1833. *The Climate of London*. Dalton: London.

Hunt M. 1997. *How Science Takes Stock: The Story of Meta-Analysis*. Russell Sage Foundation: New York.

Kratzer PA. 1937. *Das Stadtklima [The Urban Climate]. Friedr*. Vieweg und Sohn: Braunschweig.

Landsberg HE. 1970. Meteorological observations in urban areas. *Meteorological Monographs* **11**: 91–99.

Landsberg HE. 1981. *The Urban Climate*. Academic Press: New York.

Large JA. 1983. *The Foreign-Language Barrier: Problems in Scientific Communication*. Andre Deutsch: London.

Lowry WP. 1977. Empirical estimation of the urban effects on climate: A problem analysis. *Journal of Applied Meteorology* **16**: 129–135.

Nakagawa K. 1996. Recent trends of urban climatological studies in Japan, with special emphasis on the thermal environments of urban areas. *Geographical Review of Japan B* **69**: 206–224.

Oke TR. 1976. The distinction between canopy and boundary-layer urban heat islands. *Atmosphere* **14**: 269–277.
Oke TR. 1979. *Review of Urban Climatology 1973–1976*. World Meteorological Organization: Geneva. WMO Technical Note No. 169.
Oke TR. 1982. The energetic basis of the urban heat island. *Quarterly Journal of the Royal Meteorological Society* **108**: 1–24.
Oke TR. 1988. The urban energy balance. *Progress in Physical Geography* **12**: 471–508.
Oke TR. 1995. Classics in physical geography revisited – Sundborg A. 1951: Climatological studies in Uppsala with special regard to the temperature conditions in the urban area. *Progress in Physical Geography* **19**: 107–113.
Oke TR. 2004. Initial Guidance to Obtain Representative Meteorological Observations at Urban Sites. IOM Report 81. World Meteorological Organization: Geneva.
Oke TR. 2006. Towards better scientific communication in urban climate. *Theoretical and Applied Climatology* **84**: 179–190.
Oke TR. 2009. The need to establish protocols in urban heat island work. Paper presented at the *T.R. Oke Symposium & Eighth Symposium on Urban Environment*, 11–15 January, Phoenix. URL http://ams.confex.com/ams/89annual/techprogram/paper_150552.htm.
Oke TR, Maxwell GB. 1975. Urban heat island dynamics in Montreal and Vancouver. *Atmospheric Environment* **9**: 191–200.
Parry M. 1956. Local temperature variations in the Reading area. *Quarterly Journal of the Royal Meteorological Society* **82**: 45–57.
Peterson JT. 1973. The climate of cities: A survey of recent literature. In *Climate in Review*, McBoyle G (ed). Houghton Mifflin: Boston.
Pigott TD. 1994. Methods for handling missing data in research synthesis. In *The Handbook of Research Synthesis*, Cooper H, Hedges LV (eds). Russell Sage Foundation: New York.
Roth M. 2007. Review of urban climate research in (sub)tropical regions. *International Journal of Climatology* **27**: 1859–1873.
Runnalls KE, Oke TR. 2006. A technique to detect microclimatic inhomogeneities in historical records of screen-level air temperature. *Journal of Climate* **19**: 959–978.
Stewart ID. 2007. Landscape representation and the urban-rural dichotomy in empirical urban heat island literature, 1950–2006. *Acta Climatologica et Chorologica* **40–41**: 111–121.
Stewart ID, Oke TR. 2009a. Conference notebook – A new classification system for urban climate sites. *Bulletin of the American Meteorological Society* **90**: 922–923.
Stewart ID, Oke TR. 2009b. Classifying urban climate field sites by "local climate zones": The case of Nagano, Japan. In *Preprints, Seventh International Conference on Urban Climate,* 29 June–3 July, Yokohama.
Sundborg A. 1951. *Climatological studies in Uppsala with special regard to the temperature conditions in the urban area*. *Geographica 22*. Geographical Institute of Uppsala: Sweden.
Szymanowski M. 2005. Interactions between thermal advection in frontal zones and the urban heat island of Wroclaw, Poland. *Theoretical and Applied Climatology* **82**: 207–224.
Valiela I. 2001. *Doing Science: Design, Analysis, and Communication of Scientific Research*. Oxford University Press: New York.
Wanner H, Filliger P. 1989. Orographic influence on urban climate. *Weather and Climate* **9**: 22–28.
World Meteorological Organization (WMO) 1983. *Guide to Meteorological Instruments and Methods of Observation*. WMO-No. 8. World Meteorological Organization: Geneva.

Appendix A

Table A.I. Top tier studies of the heat island literature sample

Paper title	Year	Source	Data collection	MDI[a]	Score[b]	Rank[c]
Studies of the development and thermal structure of the urban boundary-layer in Uppsala	1980	Meteorol. Inst. (Uppsala)	Stationary	0	18	1
Study of the subarctic heat island at Fairbanks, Alaska	1978	Env. Prot. Agenc. (North Carolina)	Stationary	0	18	1
Urban-rural contrasts of meteorological parameters in Lodz	2006	*Theor. Appl. Climatol.*	Stationary	0	17.5	3
Pseudovertical temperature profiles and the urban heat island measured by a temperature datalogger network in Phoenix, Arizona	2005	*J. Appl. Meteor.*	Stationary	0	17.5	3
The relationship between heat island intensity and rural land coverage in Obuse, Nagano [In Japanese.]	1999	*Tenki*	Mobile	0	17	5
Relation between heat island intensity and city size indices/urban canopy characteristics in settlements of Nagano basin, Japan	2005	*Geogr. Rev. Japan*	Mobile	0	16.5	6
The urban heat island and local temperature variations in Orlando, Florida	2006	*Southeast. Geogr.*	Stationary	0	16.5	6
Temporal and spatial characteristics of the urban heat island of Lodz, Poland	1999	*Atmos. Environ.*	Stationary	0	16	8
Influence of urban morphology and sea breeze on hot humid microclimate: The case of Colombo, Sri Lanka	2006	*Clim. Res.*	Stationary	0	15.5	9
Climatological studies in Uppsala	1951	*Geographica*	Mobile	0	15.5	10
Temporal dynamics of the urban heat island of Singapore	2006	*Int. J. Climatol.*	Stationary	0	15.5	10
Urban heat island dynamics in Montreal and Vancouver	1975	*Atmos. Environ*	Mobile	0.15	15	12
Some aspects of urban micro-climate in Kuala Lumpur, West Malaysia	1972	*Akademika*	Stationary	0.15	14.5	13
The urban heat island of a city in an arid zone: The case of Eilat, Israel	2006	*Theor. Appl. Climatol.*	Mobile	0	14.5	13
Dynamics and controls of the near-surface heat island of Vancouver, British Columbia	2000	*Phys. Geogr.*	Mobile	0.15	14.5	15
Influence of meteorological conditions on the urban heat island effect in Regina	2000	*Can. Geogr.*	Stationary	0	14.5	15
Observations on the effect of a city's form and function on temperature patterns	1970	*New Zeal. Geogr.*	Mobile	0.23	13	17
Temperature and humidity traverses across London	1962	*Weather*	Mobile	0.15	13	17
Observations on the effect of a city's form and functions on temperature patterns: A case of Kuala Lumpur	1973	*J. Trop. Geogr.*	Mobile	0.15	12.5	19

[a] Missing data index. [b] Out of 18 points. [c] $N = 190$.