Understanding and assessing uncertainty of observational climate datasets for model evaluation using ensembles

Marius Zumwald¹,² | Benedikt Knüsel¹,² | Christoph Baumberger¹ | Gertrude Hirsch Hadorn¹ | David N. Bresch¹,³ | Reto Knutti²

¹Institute for Environmental Decisions, ETH Zürich, Zürich, Switzerland
²Institute for Atmospheric and Climate Science, ETH Zürich, Zürich, Switzerland
³Swiss Federal Office of Meteorology and Climatology MeteoSwiss, Zürich, Switzerland

Abstract

In climate science, observational gridded climate datasets that are based on in situ measurements serve as evidence for scientific claims and they are used to both calibrate and evaluate models. However, datasets only represent selected aspects of the real world, so when they are used for a specific purpose they can be a source of uncertainty. Here, we present a framework for understanding this uncertainty of observational datasets which distinguishes three general sources of uncertainty: (1) uncertainty that arises during the generation of the dataset; (2) uncertainty due to biased samples; and (3) uncertainty that arises due to the choice of abstract properties, such as resolution and metric. Based on this framework, we identify four different types of dataset ensembles—parametric, structural, resampling, and property ensembles—as tools to understand and assess uncertainties arising from the use of datasets for a specific purpose. We advocate for a more systematic generation of dataset ensembles by using these sorts of tools. Finally, we discuss the use of dataset ensembles in climate model evaluation. We argue that a more systematic understanding and assessment of dataset uncertainty is needed to allow for a more reliable uncertainty assessment in the context of model evaluation. The more systematic use of such a framework would be beneficial for both scientific reasoning and scientific policy advice based on climate datasets.

1 | INTRODUCTION

Observational climate datasets represent physical variables and their spatio-temporal change. They serve as evidence for scientific claims, provide boundary conditions and initial values for simulations, are used to calibrate uncertain parametrizations and to evaluate climate models by comparing model outputs to observations. However, datasets only...
represent selected aspects of the real world and are subject to uncertainty. In general terms (epistemic), uncertainty is a lack of knowledge. Representational uncertainty of a dataset is uncertainty regarding how accurately the dataset represents the phenomenon it aims to measure. Non-representational uncertainties arise because abstract properties of datasets, for example, the resolution and the baseline period in global temperature datasets, are more or less adequate for a specific purpose. These non-representational uncertainties do not arise because of the dataset relationship with reality but because of the intended purpose to which the dataset is put. In this article, the general term “dataset uncertainty” is used to refer to both representational and non-representational uncertainty. To understand the uncertainties of datasets one can aim for an end-to-end uncertainty analysis using concepts from error and uncertainty propagation or use ensembles of datasets. Ensembles vary in certain aspects of their generation and are constructed increasingly often (see Cornes, Schrier, Besselaar, & Jones, 2018; Morice, Kennedy, Rayner, & Jones, 2012).

Scientists who create and curate climate datasets are well aware of limitations giving rise to biases and uncertainties (see Kennedy (2014) for sea surface temperature datasets, see Mittaz, Merchant, and Woolliams (2019) and Povey and Grainger (2015) for remote sensing datasets) as well as of the different needs of users of datasets (see Kent et al., 2019). However, there have been few systematic intercomparison projects for datasets, analogous in size and systematicity to the coupled model intercomparison project (CMIP5) for climate models (Taylor, Stouffer, & Meehl, 2012), and dataset uncertainty is not always adequately accounted for by users of climate datasets. In particular, dataset uncertainty is not systematically included in most model evaluations, which is the specific use case of datasets that we will focus on in this article. The Intergovernmental Panel on Climate Change (IPCC) chapter on model evaluation (see Flato et al., 2013) uses only one or two different datasets per variable evaluated. There are several possible reasons for this. A practical reason might be that there are no established conceptual frameworks and quantitative methods for comparing ensembles of models to ensembles of datasets. Conceptual tools that help to link dataset uncertainties to systematic dataset ensembles are likely to contribute to the development of such tools for model-data comparisons. In this article, we provide such tools. A further reason why dataset uncertainty is often ignored by climate modelers could be that scientists naïvely or at least pragmatically think of datasets as ground truth (for a discussion see Lloyd, 2018). Dataset uncertainty may also be ignored because model uncertainty is assumed to dominate when inferring from simulation results to the real-world climate system (see Knutti et al., 2017).

However, accounting for dataset uncertainty in a more systematic manner is part of good scientific practice and of increasing importance for at least two additional reasons. First, as climate models improve, disagreements between climate model projections and observations have at times been found to be due to limitations (e.g., biases) in observational data. An example was the so-called global warming hiatus, an apparent pause of the observed global warming after the year 1998. One of the reasons for the perceived hiatus was incomplete coverage, for example, in the widely used HadCRUT4 temperature dataset (Cowtan & Way, 2014). Another reason was unaccounted biases due to changes in measurement practices of sea surface temperature records (see Hausfather et al., 2017; Karl et al., 2015; for a discussion of additional reasons, see Medhaug, Stolpe, Fischer, & Knutti, 2017). In another example, datasets of ocean temperatures did not show the temperature increase predicted by climate models. Data of better quality finally revealed that the disagreement arose due to errors in the data and not in the model results (Cheng, Abraham, Hausfather, & Trenberth, 2019; Levitus, Antonov, Boyer, & Stephens, 2000). Ignoring dataset uncertainty can thus lead to false or distorted conclusions. Second, accounting for dataset uncertainty in a more systematic way is also of increasing importance because of challenges with dataset homogeneity, due to combining different sensor types such as drifting buoys and Argo floats (Hausfather et al., 2017). Moreover, in the age of big data, low-cost sensors deployed by organizations and individuals become more ubiquitously available, leading to potentially more and more diverse datasets. This allows to measure phenomena in new ways or in higher temporal and spatial resolution at the expense of measurement standardization, which potentially increases the uncertainty of datasets used in scientific endeavors.

In this interdisciplinary paper, we investigate how and why observational climate datasets are uncertain, and we propose to use dataset ensembles to understand and assess dataset uncertainty. While a more systematic treatment and understanding of uncertainty in datasets do not automatically make users of the data consider these uncertainties, we do believe that it can contribute to a more systematic treatment of uncertainty. By reviewing literature from philosophy of science, observational climate dataset creation, and climate modeling, we show how different quantifiable and unquantifiable uncertainties can be conceptualized, understood, and assessed using the pragmatic approach of ensembles. We focus on ensembles rather than other approaches like error propagation or end-to-end uncertainty quantification for three reasons. First, ensembles can also be applied in situations where uncertainty cannot be expressed by means of a probability density function. Second, ensembles can, in principle, be used to combine different uncertainty types, including those that can, and those that cannot, be expressed by probability statements. Third, due to its
conceptual simplicity and its widespread use in climate modeling, the ensemble technique is particularly suited to establish a terminology and conceptualization that is understood by many scientists including creators and users of datasets. While focusing on climate model evaluation, we provide a general framework which aims at establishing a more common understanding of observational dataset uncertainty between dataset creators and users. Dataset providers and dataset users can use the framework as a checklist for identifying possible uncertainties arising from the use of a dataset that need to be considered. Furthermore, the construction and use of ensembles as a way to assess these uncertainties allows to consistently frame and understand uncertainty along the entire chain, from measuring, creating, using, and reusing datasets. Finally, the framework also provides the conceptual tools to achieve terminological clarity regarding dataset uncertainty, which can facilitate the communication about dataset uncertainty between dataset providers and users. However, providing specific guidance how to proceed in such assessments is beyond the scope of this article.

In section 2, we review existing uncertainty conceptualizations and frameworks and highlight the need for a framework addressing issues specifically concerning climate datasets. In section 3, we provide a framework for understanding representational and non-representational dataset uncertainties, which distinguishes between three general sources of representational uncertainty. Based on this framework, we introduce in section 4 four ensemble types that can be used to assess uncertainties. In section 5, we discuss how ensembles can be interpreted and use our framework to propose strategies for creating more systematic ensembles of datasets, which can help to resolve problems creators and users of climate datasets face. In section 6, we address challenges that arise when using dataset ensembles specifically for climate model evaluation. Second, we discuss implications of our considerations for the use of increasingly available data from low-cost sensors for climate datasets. These are deployed in large numbers by organizations and individuals with increasing digitalization and the availability of novel data transmission protocols. We conclude in section 7.

2 | EXISTING UNCERTAINTY CONCEPTS AND FRAMEWORKS

In climate science, as in other scientific disciplines, “uncertainty” is an important concept. While a consensus about terminology and classification of uncertainties is lacking to date (Frigg, Thompson, & Werndl, 2015), there are many suggestions that highlight certain aspects of uncertainty or aim at a full characterization of relevant uncertainty types or sources for a specific purpose. One of the most fundamental distinctions can be drawn between epistemic and aleatory uncertainty, which in the terminology of Walker et al. (2003) concerns the nature of uncertainty. Epistemic uncertainty results from a lack of knowledge about the system under investigation, for example, an imperfect understanding of physical processes, and can thus be reduced by more research. Aleatory uncertainty, in contrast, is a property of the system itself, for example, natural variability in the climate system.

In various sub-disciplines, specific sources of uncertainty are identified. Sources important to climate modeling are boundary and initial condition uncertainty, emission scenario uncertainty, and observational data uncertainty. However, also more general sources such as imperfect models and limited theoretical understanding are mentioned (Knutti, 2008, 2018). For climate datasets, examples of sources of uncertainties are grid-box sampling uncertainties (Kennedy, Rayner, Smith, Parker, & Saunby, 2011a) and analysis uncertainty (Kennedy, 2014), which refers to the uncertainty arising when interpolating values conditional on the available data. In the context of statistical modeling, this is sometimes also called prediction uncertainty (Fouedjio & Klump, 2019).

A common conceptualization distinguishes between parametric and structural uncertainty of either climate datasets (see, for example, Kennedy, 2014) and for climate models (see, for example, Knutti, 2008). Parametric uncertainty arises due to uncertainty of specific model parameters and structural uncertainty due to underdetermined model structures. A further common conceptualization considers the extent to which we are uncertain about something. In the terminology of Walker et al. (2003), this is called the severity or level of uncertainty. Something can be known with certainty, expressed by means of a probability statement, by means of possibilities, we can know that we are ignorant about certain aspects or we can be completely ignorant about it, so-called unknown unknowns (see Kennedy, 2014). Furthermore, it has been noted that the assessment of the level of uncertainty itself can be subject to substantial uncertainty (for this second-order uncertainty see Knüsel, Baumberger, Zumwald, Bresch, & Knutti, 2020; Parker, 2014; Steel, 2016).

For climate datasets, measurement uncertainty is of central importance. In metrology, the Guide to the Expression of Uncertainty in Measurement (GUM) provides important definitions and conceptualization of uncertainty (JCGM, 2008a). In this framework, uncertainty \( \sigma \) (of measurement) refers to any “parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the
measurand)), which describes the quantity being measured (JCGM, 2008a, p. 22). Specific sources of uncertainty listed are for instance imperfect knowledge of the effect of environmental conditions and inexact values of measurement standards. The GUM also provides guidelines to quantify and assess uncertainties. Furthermore, this framework provides guidance on how to assign sources of uncertainty to probability density functions, either by the statistical analysis of series of observations or by other means such as expert knowledge and how to combine correlated and uncorrelated uncertainties. This approach allows for end-to-end uncertainty analysis using Monte Carlo resampling methods (JCGM, 2008b). While being a consistent and powerful framework, its application is limited since many uncertainties of more complex datasets, for example, those due to the underdetermination of structural modeling choices, cannot be represented as a probability density function (for a discussion of this point for climate modeling more generally, see Parker, 2010; van der Pol & Hinkel, 2019).

A promising approach to characterize uncertainty when probability functions are not available are ensembles. Ensembles can be used to quantitatively analyze and assess different kinds of uncertainties within the same framework, such as parametric and structural uncertainties. An ensemble combines outputs that result from variations of models or parts of models used in climate modeling (Parker, 2013) or dataset creation (Cornes et al., 2018) to include uncertainties explicitly, and understand and assess them. Ensembles are also used intuitively by scientists, when they try to increase their confidence, hence reduce uncertainty, of a finding using different methods or datasets (such approaches are usually sub-summarized under the term “robustness analysis”, see Woodward, 2006).

Thus, many important distinctions, for example concerning the nature of uncertainty (aleatory versus epistemic uncertainty) or the source of uncertainty in modeling (parametric versus structural uncertainty) are shared among disciplines and sub-disciplines. There are also existing frameworks designed specifically for certain applications, such as for metrology (JCGM, 2008a) or for model-based decision support (Hirsch Hadorn, Brun, Soliva, Stenke, & Peter, 2015; Smith & Stern, 2011; Walker et al., 2003). However, there is no framework that systematically discusses uncertainties that cannot be expressed by the means of probability density functions, while still considering the specific sources of uncertainties relevant to climate datasets. As outlined in the introduction, such a framework is likely to help to better account for dataset uncertainty in specific applications such as model evaluation. Hence, in the next section, we introduce a framework that can be used to understand dataset uncertainty. It is largely based on conceptualizations and frameworks of uncertainty discussed in this section.

3 UNDERSTANDING DATASET UNCERTAINTY

Global gridded in-situ datasets are based on a sample of individual measurements that are aggregated and processed into a dataset (Figure 1). Our framework distinguishes two general sources of representational uncertainties and one source of non-representational uncertainty of datasets. First, uncertainty that arises during the generation of gridded-climate datasets (1) and, second, uncertainty that arises due to biased samples (2). These two sources concern the representational accuracy of a dataset, independent of a specific use case. Hence, dataset creators are mainly concerned with those two sources of uncertainty. The third source concerns non-representational uncertainty, which arises when choosing certain abstract properties, such as the resolution and the metric, the unit in which the dataset expresses its values (3). This choice does not directly affect the representational accuracy. However, it can affect whether the dataset is adequate for its intended purpose or (in other words) “the right tool for the job” (see Bokulich, 2018). Hence, this choice is a potential source of uncertainty too. The adequacy-for-purpose evaluation is mostly conducted by the user of a dataset under consideration. Hence, for a successful evaluation, it is necessary that information about the first two general sources of uncertainty is available and interpreted correctly. However, the example of the coverage bias described in the introduction highlights that the availability of information and tools to deal with uncertainty does not secure that they are adequately used by users of datasets.

The first general source describes how an environmental parameter is measured (1a) and how the measurement outcome is further processed (1b). The second source of uncertainty deals with when and where a measurement has been conducted, that is, how a sample of measurement represents the phenomenon of interest (2). A gridded dataset has certain properties such as a metric, which defines the unit in which the dataset expresses its values, and a resolution. An example of a metric is any specific extreme temperature index describing extremes or temperature anomaly using a certain baseline period. Such a dataset is generated using certain measurement and processing methods and is used to represent and investigate a phenomenon. However, a dataset can be used to represent different phenomena or serve as evidence for different claims and does not have a fixed representational value (Leonelli, 2019). Hence, a dataset user
needs to argue why certain properties and generation methods are adequate for the intended purpose, leading to a description of the phenomenon which is used for further investigations. This evaluation of the adequacy-for-purpose can reveal further uncertainties as it may be unclear which dataset properties are best suited for an intended purpose (3).

3.1 Uncertainty of measurement and processing procedures

We distinguish two phases in the generation of datasets, measurement, and processing, which can both lead to uncertainty. The term “measuring” describes an “empirical information-gathering activity, involving physical interaction with the entity measured” (Parker, 2017, p. 279). After the decision has been made about what to measure and how to measure it, the measurement itself has two stages. The first consists in a physical interaction between a measurement device and an aspect of the target. For instance, a resistance-based thermometer sensor stores information as bits in a memory. These bits by themselves do not represent any real-world quantity. Such an instrument indication (Tal, 2017) has no meaning in itself; it should rather be seen as a physical precondition which allows to subsequently relate an instrument indication to a target (Van Fraassen, 2008). Hence, a second step is needed to create a measurement outcome that provides information about an aspect of the target. Such measurement outcomes, quantities attributed to a target, are obtained by model-based inferences (point 1a in Figure 1) from bits stored in the memory of a device to an aspect of the target under investigation (Tal, 2013, 2017). For instance, we need calibrated models of the measurement process itself to relate bits generated by a resistance sensor to a temperature metric.

Specific sources of uncertainties in measurements include the accuracy of the sensor, for example, due to finite instrument resolution or discrimination thresholds, the quality and appropriateness of the model calibration procedure, and an incomplete definition of the entity being measured (JCGM, 2008a, p. 6). Measurement standards help to reduce these uncertainties. For instance, to guarantee comparability, the World Meteorological Organization (WMO) requires that temperature measurements be taken “over level ground, freely exposed to sunshine and wind and not shielded by, or close to, trees, buildings and other obstructions” (World Meteorological Organization, 2008, p. 66). This helps to ensure that the models employed are appropriate for the measurement conditions. Furthermore, they are required to measure at a height of 1.25–2 m above ground level. The WMO also defines so-called field standards, which are setups for a reference measurement with an 0.04 K uncertainty (at 95% confidence interval; Wylie & Lalas, 1992). These two factors of contextual standardization and reference measurements are needed to ensure that measurements are comparable over long time periods and across locations. There are many historical examples where inhomogeneities over time resulted from changes in measurement techniques. For example, sea surface temperature measurements were
historically measured by taking water samples by buckets. Today, these measurements are mostly obtained from
drifting buoys and Argo floats.

Increasingly, low-cost sensors that measure meteorological variables are deployed by individuals, private companies,
non-governmental organizations, and governmental agencies. Certain private weather station networks already include
over 200,000 stations worldwide\(^5\) (compared to about 5\,500 stations used for HadCRUT4, around 27,000 stations by
NOAA and NASA, and around 36,000 used by Berkeley Earth), and such data has already been used in scientific studies
(see Ho, Knudby, Xu, Hodul, & Aminipouri, 2016), and improved the skill of operational weather models (Nipen,
Seierstad, Lussana, Kristiansen, & Hov, 2019). However, while the opportunities of using new sources of data seem
large (see Knüsel et al., 2019), using such sensors requires an analysis in a separate paper regarding the sources and
types of uncertainties in datasets, especially of the measurement and of contextually biased samples (Meier, Fenner,
Grassmann, Otto, & Scherer, 2017; Napoly, Grassmann, Meier, & Fenner, 2018).

Three choices in the modeling of a measurement outcome are underdetermined. First, different physical measure-
ment techniques exist, for example, to measure temperature, and we might not know which one was used. Second and
third, when inferring measurement outcomes from measurement indications, parameter values and other modeling
choices are underdetermined, respectively.

Different processing steps (point 1b in Figure 1) are necessary to obtain a dataset that represents properties of a tar-
get accurately. Uncertainty arises due to underdetermination in how to process measurement outcomes onto a gridded
dataset. For instance, the value of a parameter used in a specific spatial interpolation method is underdetermined
because of limitations in our scientific knowledge, a lack of empirical data or because it has no counterpart in the real
world. In climate science, the processing steps often include homogenizations of time series (Menne & Williams, 2009;
Menne, Williams, Gleason, Rennie, & Lawrimore, 2018), interpolation, gridding and re-gridding, metric transformation,
and different kinds of bias correction (see Kennedy, Rayner, Smith, Parker, & Saunby, 2011b). Furthermore, processing
includes filter criteria for measurement outcomes in space and time, for example, thresholds of minimum station den-
sity within a grid-cell (Kennedy et al., 2011a) or thresholds for number of measurements within a time period such as
months available during a baseline anomaly period. Further uncertainty can arise because of the sequence and the kind
of generation procedures applied. For the case of temperature extremes, it has been shown that different plausible
gridding procedures may lead to different results (Dunn, Donat, & Alexander, 2014). As in the case of measurement,
the uncertainties that arise due to processing can be categorized into two fundamental different types, parametric and
structural uncertainty.

3.2 Uncertainty due to a biased sample

A gridded dataset is always based on a sample of measurements (point 2 in Figure 1), which, when not accounted for,
can introduce biases invisible in gridded interpolated datasets because measurements are mostly not distributed uni-
formly or randomly over the target. For example, some regions have fewer weather stations than others. Efforts are
made to increase spatial coverage of the measurement stations to reduce coverage bias in the final gridded dataset. For
example, the update from HadCRUT3 to HadCRUT4 saw an extended sample version that includes, for example, sta-
tions from Russia and a variety of former countries of the USSR (Jones et al., 2012). Since the station density changes
over time, spatial biases are best analyzed as a function of time. Some datasets, such as the Berkeley Earth Surface Tem-
perature (BEST) dataset, explicitly account for uncertainties due to sampling (Rohde, Muller, Jacobsen, Perlmutter, &
Mosher, 2013).

While coverage bias (Cowtan & Way, 2014) is the most obvious form of bias, a sample might also be temporally or
contextually biased. In a temporally biased sample, the measurements are not equally accurate over the measurement
period of interest, for example, due to an instrument problem during a certain period (see Thompson, Kennedy,
Wallace, & Jones, 2008). In addition, contextual bias can arise when an external confounding factor affects certain mea-
surements. For example, measurements of climatological variables are affected by non-climatic factors, such as the
degree of urbanization, soil-type, and topography (Hausfather et al., 2013; Mitchell Jr, 1953). Many such biases have
been detected in climate datasets, which are corrected to the extent that they are known. While the application of a bias
correction model reduces, or in an ideal case removes, the specific bias, new parametric and structural uncertainties dis-
cussed in the previous section are introduced. Generally, while biases reduce the representational accuracy of a dataset,
this need not be a problem in particular use cases. Whether a bias becomes relevant depends on the specific use case.
Hence, knowledge about biases has to be part of the adequacy-for-purpose evaluation (Figure 1). A further source of
uncertainty is the spatial prediction uncertainty, the uncertainty of an interpolation of a statistical model given the sample at hand (for a review of methods see Li & Heap, 2014).

3.3 | Uncertainty due to dataset properties

Uncertainty can also stem from the underdetermination of choices of certain abstract properties of the dataset, such as its metric and its spatial and temporal resolution. While the uncertainty from the previous two sources arises from inaccuracies in how the dataset represents its target, uncertainty from this third source arises because it is unclear, in a specific case, whether the abstract properties are appropriate for a given purpose. Uncertainty from this third source can hence only be evaluated in relation to a concrete scientific research question. It is a form of non-representational uncertainty.

The metric describes the unit in which the dataset expresses its values. A decision for a certain metric can be based on pragmatic considerations or be due to standardization and convention. For example, HadCRUT4 (Box 1) represents the surface temperature anomaly relative to the reference period 1961–1990 in degrees Celsius. One reason for using the reference period 1961–1990 was the high availability of observations during this period, hence a pragmatic consideration (Hawkins & Sutton, 2016). The climatological reference period of 30 years is an established convention aiming at standardization and intercomparability of scientific results. However, there are no strict scientific reasons not to take 35 or 40 years (Katzav & Parker, 2018; Lovejoy & Schertzer, 2013). The choice of a specific reference period in observations can affect inferences based on the dataset such as consistency with the ensemble spread of the CMIP5, as has been shown by Hawkins and Sutton (2016). Using different baseline periods leads to a different metric. However, if a small change in a baseline represents a different target phenomenon, such as global temperature anomaly is not entirely clear. Hence, which choice of a reference period is most appropriate for a trend estimate is a source of uncertainty.

A further source of uncertainty in observational datasets is the choice of a spatial and temporal resolution of a dataset. For example, spatial correlation of temperature and hence variation differs largely dependent on the spatial resolution. Hence, the choice of the resolution affects important characteristics of the dataset, such as spatial correlation or variance across and within grid-cells that directly affect inferences based on a dataset. However, choosing an adequate resolution is often restricted by pragmatic considerations such as availability of measurements.

BOX 1  The HadCRUT4 dataset as an example of assessing several specific parametric uncertainties using an ensemble approach

HadCRUT4 is a gridded, station-based parametric dataset ensemble with 100 members that represents global land and sea surface temperature anomalies relative to 1961–1990 (Morice et al., 2012). HadCRUT4 is a combinations of a land surface temperature dataset CRUTEM4 (Jones et al., 2012) and a sea surface dataset HadSST3 (Kennedy et al., 2011a, 2011b) and is based on approximately 5,500 measurement stations distributed worldwide and contains monthly data for the years 1850 until present at a 5° x 5° resolution. Uncertainties due to underdetermination of the parameter values used in the construction are accounted for by constructing a 100 ensemble members and combining them by a Monte–Carlo approach. Uncertainties considered in the land component ensemble are:

Uncertainty due to station homogenization: Uncertainty arising from a homogenization, that is, from identifying and removing artifacts in station records such as those caused by changes in measurement equipment.

Station climatological normal uncertainty: Uncertainty arising from calculating the monthly climatological averages over the reference period from 1961 to 1990.

Uncertainty due to urbanization bias correction: Uncertainty arising from the warming effect of anthropogenic land-use changes in urban environments.

Uncertainty due to exposure bias correction: Uncertainty from the use of different types of measurement sensor enclosures that have been employed over time and have introduced systematic measurement biases.
4 | FOUR TYPES OF DATASET ENSEMBLES

The framework introduced above helps to understand uncertainty of datasets. Here we focus on the use of ensembles as means to assess uncertainty of datasets. To assess it, the elements that lead to uncertainty can be systematically varied to create dataset ensembles (see Figure 2). These variations concern (i) the parameter values, (ii) the model structures, (iii) the measurement sample, and (iv) the dataset properties. Varying one of these aspects leads to parametric, structural, resampling, and property ensembles, respectively. Both, parametric and structural ensembles can be used to assess uncertainties arising due to underdetermination of measurement and processing procedures. Resampling ensembles pass subsamples of measurement outcomes into processing procedures. Property ensembles are created by varying the properties such as metric and resolution, which are all deemed adequate for the purpose at hand, given that the dataset describes the target phenomenon in a sufficiently accurate way for the intended purpose.

A dataset ensemble does not per se need to be the result of an individual variation of, for example, a parameter. Some dataset ensembles are created by drawing a sample from a posterior distribution of a statistical model conditional on available measurements—hence, in such cases, the model is inherently probabilistic. Such ensembles allow to assess uncertainty arising at several locations of uncertainty in one step. A further advantage is that “double counting” of noise can be prevented which can arise when drawing ensembles members from fields of uncertainty estimates. Examples of such ensembles are widespread in climate research and can be found in Cornes et al. (2018), Frei and Isotta (2019), and Song, Kwon, and Lee (2015). The kind of uncertainty accounted for differs. Frei and Isotta (2019) incorporated spatial prediction uncertainty, uncertainty due to using statistical methods to interpolate values in space, and model parameter uncertainty. Other approaches only draw on realizations from a fixed parameter set (Cornes et al., 2018). Also, measurement error and uncertainty could potentially be included as discussed by Frei and Isotta (2019). However, in cases where also structural elements are varied, a pure probabilistic representation is not possible anymore.

We note here that there are cases where the ensemble approach might not be feasible or appropriate and other approaches might be more suitable to understand and assess uncertainty. For example, an ensemble for analyzing extreme events would require a high temporal resolution. Such an ensemble might be prohibitively expensive from a computational perspective, because of, for example, limited storage size or computational power.

As we will see, ensemble approaches to assess dataset uncertainty are already employed. We note here that variation of uncertain elements to create ensembles only helps to assess uncertainty from sources that are known to be underdetermined or biased. This approach is hence only applicable to the extent that scientists know about the uncertainties in dataset construction. Unknown unknowns, uncertainties of whose existence we are not aware but somehow have an effect on our analysis, cannot be assessed in this way.

FIGURE 2 Variations at different locations of the framework lead to a dataset ensemble that helps to assess different sources of uncertainty. These variations concern (i) the parameter values, (ii) the model structures, (iii) the measurement sample, and (iv) the dataset properties. The choices made need to be evaluated in terms of their representational accuracy if they concern the first three points and overall adequacy-for-purpose, which concerns all four points.
4.1 Parametric ensembles

Analogous to perturbed-physics ensemble (PPE) of climate models (Box 2), parameters can be varied systematically if plausible bounds can be defined (see point i in Figure 2). Parametric uncertainty arises when the values of parameters in the construction of a dataset—that is, in measuring or processing the measurement outcomes further—are not well constrained, for example, by empirical evidence or background knowledge. In climate science, parametric dataset ensembles are increasingly generated and used. The HadCRUT4 temperature dataset (Box 1) is an example of a parametric ensemble which combines different parameter values in the processing of the dataset. Many other examples of this type of ensemble exist for different climate variables. Such ensembles are generated by varying the values of parameters that capture uncertainty due to station homogenization or uncertainty that arises due to the use of different measurement sensor enclosures. They have been used to account for uncertainties due to changes of sea surface temperature measurement conditions over time, which can be done explicitly in a parametric ensemble (see Kennedy, Rayner, Atkinson, & Killick, 2019). Another example is the pairwise homogenization algorithm (Menne & Williams, 2009) which is used to detect and correct inhomogeneities in the time series. Consequently, variations in parameters of this pairwise homogenization algorithm can then be used to account for uncertainties due to homogenization (Williams, Menne, & Thorne, 2012). For the extended reconstructed sea surface temperature dataset version 4, the combined parameter uncertainty of 24 parameters has been assessed using an ensemble of a thousand members (Huang et al., 2015). Further examples are the HadEX2 (Donat et al., 2013) and E-OBS (Cornes et al., 2018) parametric dataset ensembles, which are datasets of global temperature and precipitation extremes and Europe-wide temperature (daily minimum, mean, and maximum value) and precipitation, respectively.

**BOX 2 Climate model ensembles**

**Perturbed-physics ensemble (PPE):** An ensemble that consists of different versions of the same model structure by applying different, plausible parameter values (Knutti et al., 2008; Stainforth et al., 2005). Hence, a PPE can be viewed as an ensemble that contains instances of the same model structure representing a target.

**Multi-model ensemble (MME):** An ensemble that combines structurally different models. The models can, for instance, differ in complexity, the components included, the types of a parametrization, and other aspects. Mostly, they are developed at different institutions (see Parker, 2013).

**CMIP:** The most comprehensive completed MME to date is the fifth coupled model intercomparison project (CMIP5), which was used in the fifth assessment report of the Intergovernmental Panel on Climate Change (IPCC) (Pachauri et al., 2014) in order to understand structural model uncertainties (Flato et al., 2013; Knutti & Sediček, 2013). CMIP6 (Eyring et al., 2016) is currently being generated, which will contribute to the sixth Assessment Report of the IPCC.

**Ensemble of opportunity:** An ensemble that is not systematically created. Instead, it combines different existing and structurally different climate models and is used in a pragmatic manner to assess uncertainty (Tebaldi & Knutti, 2007).

**Model democracy:** Gives equal weight to each model when aggregating an MME. This approach assumes that all models in an ensemble are independent and equally plausible. Hence, it is assumed that the mean of all ensemble members is an accurate description of reality once variability is suppressed (see Knutti, Baumberger, & Hirsch Hadorn, 2019) However, since models are not independent realizations, there are good reasons to partly reject these assumptions (Knutti, Masson, & Gettelman, 2013).

**Model weighting:** To overcome the limitations imposed by model democracy, recent studies have developed weighting schemes in order to account for skill of individual models and model interdependence based on model output. The basic idea is that models receive more weight if they agree well with observational data and if they are independent of other models (Knutti et al., 2017; Sanderson, Wehner, & Knutti, 2017).
4.2 | Structural ensembles

Analogous to multi-model ensembles (MME) for climate models (see Box 2), a structural ensemble of datasets helps to assess structural uncertainties (see point ii in Figure 2). Structural uncertainty arises due to underdetermination of modeling choices when constructing measurement devices or when processing measurement outcomes. Datasets generated by structurally different modeling approaches have also been used as an ensemble but in a rather ad-hoc manner. Examples are ensembles of different rain-gauge measurement-based datasets (Prein & Gobiet, 2016), and ensembles used to quantify uncertainties of tropospheric temperature trends from radiosondes (Thorne et al., 2011). Medhaug et al. (2017) used datasets that are generated in structurally different ways to investigate global warming trends. While such dataset ensembles are similarly ad-hoc as the CMIP5 (Box 2) in terms of their generation, they mostly lack the broad efforts put into CMIP5. In contrary to the examples explained above, CMIP5 (Taylor et al., 2012) consists of standard experiments that assess the model response to standardized forcing scenarios (see Thomson et al., 2011) and uses standardized evaluation procedures on pre-defined metrics. Also, documentation, data format, and data exchange are standardized, which allows to more easily interpret different models as an ensemble. The Observations for Model Intercomparisons Project (Obs4MIPs) is a structural dataset ensemble of opportunity, which aims to provide an intercomparison of existing structurally different datasets to be used for model evaluation (see Ferraro, Waliser, Gleckler, Taylor, & Eyring, 2015; Teixeira et al., 2014). Obs4MIP is specifically tailored to be used in model evaluation and might not be readily applicable in other contexts. Another example is the State of the Climate report (see Blunden & Arndt, 2019), an annually published summary of the global climate, which considers existing structurally different datasets for a variety of climate variables.

4.3 | Resampling-based ensembles

Datasets based on in situ measurements use a measurement sample that is typically unequally distributed over the target. For constructing a dataset such as HadCRUT4, one needs to rely on what one might call a measurement sample of opportunity that uses data from already existing measurement stations. HadCRUT4, for example, is based on approximately 5,500 individual stations which are unequally distributed across the surface of the earth.

Methods based on resampling can be used for two reasons. First, random resampling strategies can be used to generally assess uncertainties arising because of an imperfect sample, by assessing the sensitivity of the outcome by reproducing the analysis on subsamples. In the BEST dataset (Rohde et al., 2013), statistical uncertainties are calculated by subdividing the data and comparing the results from statistically independent subsamples using the jackknife algorithm (Efron & Stein, 1981), a resampling method using a “leave one observation out at a time” approach. Bootstrapping is another approach based on sampling with replacement (Efron & Tibshirani, 1994).

Second, non-random resampling, conditional resampling based on, for example, a confounding variable, can be used to assess the effect of a contextual bias, for example, by including only stations that are affected by similar measurement contexts, which can help to assess the effect of contextual biases such as urbanization. This, however, requires (local) background knowledge about the context of measurements. To assess coverage biases by resampling methods some reference data needs to be available. Cowtan and Way (2014) investigated the coverage bias of the HadCRUT4 by filling in the temperature field with samples from satellite data. They performed this reconstruction using different subsamples of the available observational record. This can be seen as a resampling-based dataset ensemble that was created to investigate coverage biases. Generally, resampling can only help to detect biases, but not to correct them.

4.4 | Property ensembles

A property ensemble (see point iv in Figure 2) varies abstract properties of the datasets, such as the resolution and the metric. They can be used to investigate the robustness of a specific inference drawn from the dataset with respect to the property choices. For example, if a scientist decides to use the HadCRUT4 dataset to investigate whether the earth has become warmer on a global level, she needs to argue why an anomaly-based metric and a 5°C resolution might be well suited to do so. However, in many cases, there is uncertainty about the range of adequate properties such as resolution and also more than one metric can be adequate.
One can restrict the set of potentially suitable properties by using background knowledge, since our background knowledge is always closely related to a metric and how we measure things. For example, the phenomenon of global temperature anomaly can be operationalized using different reference time periods of different lengths and with different starting dates. If a climate dataset is used to investigate the claim that the global temperature anomaly has increased in the last decades, then the sensitivity of the temperature trend to the choice of the reference period allows one to investigate the robustness of this claim with respect to one dataset property. This allows to investigate the uncertainty of the decision for certain properties that are all deemed adequate for the phenomenon at hand.

Property ensembles differ from the other ensemble types in one important aspect. A dataset with certain properties can be obtained by multiple generation procedures. For example, a dataset with a certain resolution can be generated using different gridding and interpolation procedures. Obtaining a change in the properties requires a change in the generation procedures. For example, changes in the resolution of a dataset require different parameter values that lead to this dataset. Hence, changing the resolution of a dataset requires that at least some of the preceding procedures are modified, too. Thus, this has important consequences because any variation that leads to different properties also requires an assessment of the adequacy of the other changes. They are also different from the other types of ensembles because the choice of what is an adequate property is not about making choices that accurately reflect reality, but it is about choosing properties that are useful for the desired application. Using ensembles of datasets with different properties can reveal whether the result of an analysis is robust with respect to the choice of these abstract properties. In the creation of the sea surface temperature dataset HadSST2 (Rayner et al., 2006), the generation of versions with different resolution is implemented by design, which allows to more systematically create property ensembles.

5 | CONNECTING DATASET ENSEMBLE CREATORS AND USERS

To mainstream the creation and correct use of dataset ensembles, they need to be created more systematically, inter-compared, and it needs to be well documented what sources of uncertainties are accounted for in the ensemble so that dataset users understand this information. Hence, a common terminology concerning ensemble types and the specific sources included is needed to reduce misunderstandings and facilitate the correct use of dataset ensembles that are created by climate data scientists. Our framework can provide guidelines for a common terminology and conceptualization of uncertainties and ensemble types included in a dataset ensemble. This would also facilitate iterative procedures in creating datasets as information about, for example, a detected bias that feeds back into novel dataset processing procedures, leading to more accurate datasets suitable for more use cases while being explicit about uncertainties.

5.1 | Systematic dataset ensemble creation and intercomparison

Systematicity has two dimensions, which both come in degrees. First, we can be systematic in terms of how the ensemble samples the space of potential variations. Second, we can be systematic about how we intercompare and disseminate existing ensembles. Referring to the first point, an ensemble consisting of several, ideally independently created, members can be used in a rather ad-hoc fashion to test the robustness of a scientific claim by using the ensemble as an ensemble of opportunity (Box 2). However, an ideal ensemble created by using infinite resources and by considering all that is known about the target system and the elements relevant to the generation of the members would sample the uncertainty more systematically. The model intercomparison project CMIP5 lies in between being completely ad-hoc and fully systematic. Even though in CMIP5, many of the climate models are not, to a relevant degree, developed independently, the model runs, hence the creation, is highly standardized (Taylor et al., 2012). Furthermore, the intercomparison and dissemination of CMIP5 is to a relevant degree systematic. A common understanding of what set of climate models, for example, the CMIP5 ensembles consists of, facilitates comparing results between studies using this ensemble. Hence, the criteria of which member to include and which not is important here. Also in the construction of datasets, different approaches exist that are independent to a relevant degree and allow to assess uncertainty in the representation of a target phenomenon, for example, global temperature anomaly (Hansen, Ruedy, Sato, & Lo, 2010; Morice et al., 2012; Muller et al., 2014).

Since dataset ensembles of opportunity such as Obs4MIPs (see Ferraro et al., 2015; Teixeira et al., 2014) are not created systematically, they do not sample the uncertainty systematically. Theoretically, systematic variation of under-determined methodological choices involved in creating a dataset would lead to a large dataset ensemble spanning the
total representational uncertainty. However, given the limits of our current scientific knowledge, this seems rather difficult with a large number of structurally different ensembles.7 Besides those epistemic limits, there are also practical limitations, since our financial and computational resources are limited. The computational argument mainly refers to uncertainties that arise due to processing, whereas the financial argument refers mainly to measurement infrastructure, processing efforts, and people developing the scientific tools to implement the necessary steps. Ideally, scientists could draw from a dense network of in-situ measurement stations to experiment with different measurement approaches to create an ensemble of measurement and resampling processes. For instance, HadCRUT4 is built on around 5,500 measurement stations and the BEST dataset on around 36,000 stations, which already creates greater possibilities for systematic resampling. For economic reasons, however, scientists in practice depend on existing measurement infrastructures maintained by national weather organizations.

While the aforementioned practical and epistemic limits are barriers for the creation of systematic dataset ensembles, the creation of more systematic ensembles compared to those currently available could be achieved by exchanging different existing processing procedures that are used to solve the same task. This may be practical because the creation of datasets is computationally less expensive compared to running the newest generation of climate models used in CMIP6. Hence, structurally different methods and parameter ranges used for homogenization, gridding, or interpolation could be recombined systematically, leading to ensembles of opportunity concerning their processing procedures rather than the final datasets. This would require that the interfaces of the individual processing procedures be standardized, which would further improve the reliability of uncertainty estimates, reproducibility, and facilitate the identification of results that are robust to representational dataset uncertainties (see Thorne et al., 2011). Such a plug-and-play approach using different modules would not lead to any new structural elements per se but would maximize the utility of existing ones. Publishing open source software and more modular information technology architecture are additional measures toward more systematic intercomparison. While being fruitful, the analogy between datasets and climate models should be drawn with care. For datasets, convergence is probably more likely to be a sign of representing reality accurately than for climate models. Nevertheless, it cannot be excluded that convergence of datasets can still at least partly be a sign of a lacking independence (e.g., an ensemble version using the same subsample of measurement stations).

5.2 | Interpreting the ensemble spread

Using dataset ensembles to assess uncertainty requires two things. First, one needs to understand what specific sources of uncertainty are included in a dataset ensemble. Even if the parametric dataset ensemble HadCRUT4 (Box 2) is well documented, due to a lack of standardization in terminology and best practices, it is not trivial to understand what kind of uncertainties are represented. Second, the ensemble spread might be interpreted differently by different scientists in different contexts. Hence, the assumptions underlying the interpretation of what the ensemble spread represents in a statistical sense need to be stated clearly. This requires that all information regarding uncertainty is well-documented, for example, by accompanying meta-data, and is understood by scientists who use the dataset ensemble. Climate data scientists are mostly well aware of the uncertainty and limitations of their data, but often this information is lost when the data is used by modelers and other users (Brönnimann & Wintzer, 2018), van der Pol and Hinkel (2019) show that in the case of sea-level rise, the information about the uncertainty estimation by an ensemble often changes (as in a telephone game). Information about what kind of uncertainty is assessed by an ensemble is lost or unwittingly changed by imposing arbitrary or wrong assumptions (e.g., that the uncertainty can be represented by a probability density function). While Van der Pol and Hinkel make this point regarding sea level rise projections, it seems to translate to observational datasets. Moreover, current dataset ensembles are mostly of parametric nature, whereas many model ensembles in climate science are structural ensembles. This might lead to confusion and potential users of data ensembles might not understand the ensemble type and how this impacts their interpretation, since they have experience in working with a different ensemble type. Some ensembles may even combine different ensemble types such as variation in parameter values and in structural elements, which further complicates uncertainty assessment. This is because entangling the individual contribution to the uncertainty might require a separate sensitivity assessment and it is unclear how a spread of a mixed type ensemble can be interpreted.

Within a quantitative uncertainty assessment framework, the spread of a parametric dataset ensemble can be interpreted as a probability density function if certain conditions hold. Namely, the plausible range of parameter values needs to be specified probabilistically, too. Such a range can, for example, be specified in a Bayesian framework with
priors for the parameter values. Subsequently, the distribution, empirical quantiles, or skewness measures can be used to assess how the dataset changes in response to a change in a parameter. However, this requires that, as input, a plausible range of parameter values can be identified. This can be challenging in practice. Ensembles resulting from a plausible range of parameter values can be obtained by, for example, Monte–Carlo simulations, but this requires efficient recombination strategies to manage the large number of combinations.

Note that undetected or wrongly estimated biases cannot be investigated using such methods. In cases where parametric uncertainty is analyzed in isolation and can be interpreted as a probability density function, providing the standard deviations can suffice, and providing the uncertainty estimates as a collection of members is not necessary. Sometimes parameters cannot be defined as a probability density function and relate more to practical design choice blurring the boundary between parametric and structural elements.

If a large number of structurally different ensemble members were created, and if we could assure that they were completely independent of each other and equally plausible, then a structural ensemble, too, could be interpreted probabilistically. As it is unclear what the full range of adequate measurement model structures is (and if such a range even exists), a probabilistic interpretation of structural ensembles is difficult. This is widely recognized for MMEs of climate models (Parker, 2010). As is the case for climate models (Knutti et al., 2013), different datasets that are generated by structurally different modeling approaches will very likely show interdependence in some aspects, because similar measurement devices or interpolation schemes are used that are biased in the same way. In any case, a reliable uncertainty assessment requires knowledge of how to distinguish between plausible and implausible measurement model structures. Thus, in practice, such ensembles cannot be interpreted probabilistically. Structural ensembles (for MMEs see Knutti & Sedlāček, 2013) can be used to assess derivational robustness (Woodward, 2006). This allows to assess the sensitivity of the overall results to changes in certain assumptions, such as structural elements in the dataset generation, on the outcome. Also, property, resampling, and parametric ensembles should always be used to investigate the derivational robustness of variations in the properties or subsample.

6 USING DATASET ENSEMBLES IN CLIMATE MODEL EVALUATION

In this section, we discuss how to make use of dataset ensembles in climate model evaluation, for example, when developing standardized methods on how to compare dataset ensembles of different types with climate model ensembles. This is especially important to make use of information regarding uncertainty encoded in dataset ensembles, which are increasingly being constructed. The formulation of needs of users that need to be understood by creators depends on an understanding of the methodological and technical possibilities of creating ensembles. Hence, as discussed in section 5, this is something that requires constant exchange between dataset creators and users. Using the full information of climate model ensembles is standard in climate science, and also climate impact studies mostly use more than one climate model (Burke, Dykema, Lobell, Miguel, & Satyanath, 2015). However, when evaluating climate model ensembles, the uncertainty of datasets is often not considered. For example, in CMIPS, climate models were evaluated by calculating a distance metric, such as space–time root-mean-square error (RMSE), with respect to only one or two datasets for different variables (Flato et al., 2013).

To reduce the complexity for decision-making or to approximate the real value, climate model ensembles are aggregated to one projection. Recently, it was suggested to weight individual climate models in ensemble studies with respect to their independence and skill, measured as distance to a dataset (Knutti et al., 2017). Brunner, Lorenz, Zumwald, and Knutti (2019) show that including dataset uncertainty when weighting climate models increases the robustness of the results, independent of the chosen approach to include dataset uncertainty. However, the approaches presented by Brunner et al. (2019) assume that all datasets are independently constructed. This independence assumption seems implausible both for datasets (as has been discussed above) and for climate models (see Knutti et al., 2013).

Further research is needed to better understand the interdependence between datasets. One needs to separate between measures of interdependence based on the metric of the dataset itself, interdependence of the methods used to generate a dataset, and interdependence due to shared measurements. For the first type of interdependence, one could aggregate datasets purely based on a typical evaluation metric, for example, space–time RMSE between dataset ensemble members. However, it needs to be empirically investigated whether this is an adequate measure of interdependence as has already been done for climate models (see Knutti et al., 2013). If empirical investigations reveal that an independence weighting based on a space–time distance metric cannot be applied, methods and frameworks based on structured expert elicitation, as it has been suggested for uncertainty quantification of climate models (Oppenheimer,
Little, & Cooke, 2016), can be used to define the independence of the methods used to generate different datasets. Independent measurement stations and procedures are in practice not very common, since most gridded dataset rely on a contingent sample of measurement stations. Regardless of the method used and type of independence, an independence-weighted dataset could subsequently be used as input into already developed weighting schemes (Brunner et al., 2019).

Reanalysis datasets such as ERA5 (Hersbach et al., 2019) are often used in model evaluation because their complete global fields allow for an easy comparison with model outputs. These datasets are obtained by using climate models in a data assimilation procedure to create globally complete datasets. It is still debated how the reliability of the information from reanalysis datasets compares to that in measurement-based datasets (Parker, 2016). We do not consider reanalysis datasets as strictly observational datasets and have hence not discussed them further. Nevertheless, we think that our framework can, in principle, also be applied to understand the uncertainty of reanalysis datasets.

7 | CONCLUSIONS

Accounting for uncertainty in climate datasets is of increasing relevance since the skill of climate models is increasing and their use for regional and local decision making is gaining importance. A more systematic understanding and assessment of dataset uncertainty allows for a more reliable assessment of total uncertainty, beneficial both to scientific reasoning and to scientific policy advice.

Our framework allows to identify and classify from where uncertainty arises when generating and using datasets. It shows how ensembles of different types can in principle be used to assess different general sources of uncertainties. To include dataset uncertainty in scientific analyses, both the dataset generation and the climate modeling communities are required to exchange their needs and challenges such that the relevant uncertainties can be considered. A shared terminology and conceptualization about specific and general sources of uncertainties are of importance to mitigate misconceptions when dataset ensembles are used by various scientists in diverse contexts. It also facilitates discussion about desirable properties of a dataset ensemble for specific purposes because such ensembles are always created under resource constraints. Including dataset uncertainty into model evaluation and other analyses is important. When aggregating and weighting climate model ensembles, there is no standardized way to include observational dataset ensembles. Further research is needed to improve climate model evaluation using such dataset ensembles. Finally, further work is needed to show how to separate the different uncertainties that are assessed by different ensemble types and how to communicate them more clearly to decision makers.

While the focus of our review is on observational climate datasets and climate model evaluation, we assume that our conceptualization provides insights into dataset uncertainty more generally. However, further research is needed to show whether the proposed framework and tools are appropriate for other datasets and uses.

ACKNOWLEDGMENTS

We thank Stefan Brönnimann, Sven Ove Hansson, and Wendy Parker for feedback on an earlier version of this paper. We are furthermore grateful for the valuable comments of two anonymous reviewers. The research was supported by the Swiss National Science Foundation, National Research Programme 75 Big Data, project number 167215.

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

AUTHOR CONTRIBUTIONS

Marius Zumwald: Conceptualization; writing-original draft; writing-review and editing. Benedikt Knüsel: Conceptualization; writing-original draft; writing-review and editing. Christoph Baumberger: Conceptualization; writing-original draft; writing-review and editing. Gertrude Hirsch Hadorn: Conceptualization; writing-original draft; writing-review and editing. David Bresch: Conceptualization; writing-original draft; writing-review and editing. Reto Knutti: Conceptualization; writing-original draft; writing-review and editing.

ORCID

Marius Zumwald https://orcid.org/0000-0003-4763-2374
Benedikt Knüsel https://orcid.org/0000-0003-0292-5982
ENDNOTES

1 Measurement uncertainty should not be confused with measurement error. Imperfections in measurement give rise to an error whereas measurement uncertainty reflects the lack of knowledge of the exact true value of the measurand, the object being measured, which affects our assessment of the error. Consequently the measurement error can be small (when the measurement outcome is close to the true value) despite having a large uncertainty (JCGM, 2008a, p.6). Nevertheless, we note here that measurement errors that cannot be corrected can be a source of measurement uncertainty.

2 Our use of the term “metric” differs from the use of the term “measurand” in metrology, which describes “a well-defined physical quantity that can be characterized by an essentially unique value” (JCGM, 2008a, p. 1) or “a particular quantity subject to measurement” (JCGM, 2008a, p. 34). An example is the “Vapor pressure of a given sample of water at 20°C” (JCGM, 2008a, p. 34). Hence, the measurand consists of a metric (i.e., vapor pressure at 20°C) attributed to a particular object in the real-world (i.e., given sample of water). Metrology mainly deals with measuring quantities that are rather closely derived from the seven basic SI-metrics which are well-defined standards that are clearly relatable to our physical world. However, many datasets use more complex metrics, such as a temperature extremes index or temperature anomaly using a specific baseline period. Although they can be derived from SI-metrics, it is much less clear what the physical quantity, the real-world object, corresponding to for example, a temperature extreme index is. Since a metric is a property of the dataset, any scientist needs to argue whether this property is adequate for a specific purpose, which corresponds to the third general source of uncertainty in our framework.

3 For the case of in-situ based temperature datasets, we use the term “measurement” for measuring temperature at a certain spatial and temporal location using specific physical devices and theories. All later activities we call “processing” of the measurement outcomes. We are aware that the distinction between modeling a measurement outcome and further processing the outcome can be drawn in different ways. However, there are two reasons to draw the distinction as suggested. First, a temperature value of a measurement device is the most unprocessed information that establishes a relationship to the target and can be manipulated by the scientists. Second, the calibration of measurement devices is a well-established practice which is done by people other than providers of climate datasets. The suggested distinction is well in line with how the terms “measurement” and “processing” are used in climate science.

4 Besides using electrical resistance elements, thermoelectric sensors can also be based on semiconductors or thermocouples.

5 https://www.wunderground.com/blog/JeffMasters/wu-personal-weather-stations-are-now-200000-strong.html.

6 We note here that the way in which an ensemble is created and presented also affects how it is used and how the included uncertainties are understood by users. For instance, an ensemble can include a best estimate, come as ordered collection or in a pure random order. Parameter values in a parametric ensemble can be chosen randomly or be predefined. Hence, the interaction between the construction and presentation of an ensemble, on the one hand, and its interpretation on the other hand needs to be better understood. However, such points are beyond the scope of our manuscript and are not discussed further.

7 As mentioned in section 3, we do not discuss uncertainty that arises because of recognized or total ignorance. Assessing such uncertainties requires other strategies than ensembles.

RELATED WIREs ARTICLES
Skill and uncertainty in climate models
Ensemble modeling, uncertainty and robust predictions
Quantifying the irreducible uncertainty in near-term climate projections

REFERENCES
Blunden, J., & Arndt, D. S. (2019). State of the climate in 2018. Bulletin of the American Meteorological Society, 100(9), S1–S306. https://doi.org/10.1175/2019BAMSStateoftheClimate.1

Bokulich, A. (2018). Using models to correct data: paleodiversity and the fossil record. Synthese. https://doi.org/10.1007/s11229-018-1820-x
Thomson, A. M., Calvin, K. V., Smith, S. J., Kyle, G. P., Volke, A., Patel, P., ... Edmonds, J. A. (2011). RCP4.5: A pathway for stabilization of radiative forcing by 2100. *Climatic Change*, 109(1), 77. https://doi.org/10.1007/s10584-011-0151-4

Thorne, P. W., Brohan, P., Titchner, H. A., McCarthy, M. P., Sherwood, S. C., Peterson, T. C., ... Kennedy, J. J. (2011). A quantification of uncertainties in historical tropical tropospheric temperature trends from radiosondes. *Journal of Geophysical Research-Atmospheres*, 116(D12), 1–19. https://doi.org/10.1029/2010JD015487

van der Pol, T. D., & Hinkel, J. (2019). Uncertainty representations of mean sea-level change: A telephone game? *Climatic Change*, 152(3), 393–411. https://doi.org/10.1007/s10584-018-2359-z

van Fraassen, B. C. (2008). *Scientific representation: Paradoxes of perspective*. Oxford: Clarendon Press/Oxford University Press.

Walker, W. E., Harremoës, P., Rotmans, J., van der Sluijs, J. P., van Asselt, M. B. A., Janssen, P., & von Krauss, M. P. K. (2003). Defining uncertainty: A conceptual basis for uncertainty Management in Model-Based Decision Support. *Integrated Assessment*, 4(1), 5–17. https://doi.org/10.1076/iaij.4.1.5.16466

Williams, C. N., Menne, M. J., & Thorne, P. W. (2012). Benchmarking the performance of pairwise homogenization of surface temperatures in the United States. *Journal of Geophysical Research-Atmospheres*, 117(D5), 1–16. https://doi.org/10.1029/2011JD016761

Woodward, J. (2006). Some varieties of robustness. *Journal of Economic Methodology*, 13(2), 219–240. https://doi.org/10.1080/13501780600733376

World Meteorological Organization. (2008). *Guide to meteorological instruments and methods of observation*. Geneva, Switzerland: World Meteorological Organization.

Wylie, R. G., & Lalas, T. (1992). *Measurement of temperature and humidity: Specification, construction, properties and use of the WMO reference psychrometer*. Geneva, Switzerland: World Meteorological Organization.

**How to cite this article:** Zumwald M, Knüsel B, Baumberger C, Hirsch Hadorn G, Bresch DN, Knutti R. Understanding and assessing uncertainty of observational climate datasets for model evaluation using ensembles. *WIREs Clim Change*. 2020;11:e654. https://doi.org/10.1002/wcc.654