Received: 28 February 2019 | Revised: 05 June 2019 | Accepted: 21 June 2019 | Published on: 23 August 2019

DOI: 10.1002/qj.3598

INVITED MANUSCRIPT

Quarterly Journal of the Royal Meteorological Society RMetS

Towards a more reliable historical reanalysis: Improvements for version 3 of the Twentieth Century Reanalysis system

Laura C. Slivinski[1,2] | Gilbert P. Compo[1,2] | Jeffrey S. Whitaker[2] | Prashant D. Sardeshmukh[1,2] | Benjamin S. Giese[3] | Chesley McColl[1,2] | Rob Allan[4] | Xungang Yin[5,6] | Russell Vose[6] | Holly Titchner[4] | John Kennedy[4] | Lawrence J. Spencer[1,2] | Linden Ashcroft[7] | Stefan Brönnimann[8] | Manola Brunet[9,10] | Dario Camuffo[11] | Richard Cornes[12] | Thomas A. Cram[13] | Richard Crouthamel[14] | Fernando Domínguez-Castro[15] | J. Eric Freeman[5,6] | Joëlle Gergis[16,17] | Ed Hawkins[18] | Philip D. Jones[10] | Sylvie Jourdain[19] | Alexey Kaplan[20] | Hisayuki Kubota[21] | Frank Le Blancq[22] | Tsz-Cheung Lee[23] | Andrew Lorrey[24] | Jürg Luterbacher[25] | Maurizio Maugeri[26] | Cary J. Mock[27] | G.W. Kent Moore[28] | Rajmund Przybylak[29] | Christa Pudmenzky[30] | Chris Reason[31] | Victoria C. Slonosky[32] | Catherine A. Smith[1,2] | Birger Tinz[33] | Blair Trewin[7] | Maria Antónia Valente[34] | Xiaolan L. Wang[35] | Clive Wilkinson[10] | Kevin Wood[36,37] | Przemysław Wyszyński[29]

[1]Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado, USA
[2]Physical Sciences Division, NOAA Earth System Research Laboratory, Boulder, Colorado, USA
[3]Texas A&M University, College Station, Texas, USA
[4]Hadley Centre, Met Office, Exeter, UK
[5]Riverside Technology Inc, Asheville, North Carolina, USA
[6]National Centers for Environmental Information, Asheville, North Carolina, USA
[7]School of Earth Sciences, University of Melbourne, Australia
[8]Oeschger Centre for Climate Change Research and Institute of Geography, University of Bern, Switzerland
[9]Centre for Climate Change, Universitat Rovira i Virgili, Tarragona, Spain
[10]Climatic Research Unit, University of East Anglia, Norwich, UK
[11]National Research Council, Institute of Atmospheric Sciences and Climate, Padua, Italy
[12]National Oceanography Centre, Southampton, UK
[13]National Center for Atmospheric Research, Boulder, Colorado, USA
[14]International Environmental Data Rescue Organization, Deale, Maryland, USA
[15]Instituto Pirenaico de Ecología, Zaragoza, Spain

Abstract

Historical reanalyses that span more than a century are needed for a wide range of studies, from understanding large-scale climate trends to diagnosing the impacts of individual historical extreme weather events. The Twentieth Century Reanalysis (20CR) Project is an effort to fill this need. It is supported by the National Oceanic and Atmospheric Administration (NOAA), the Cooperative Institute for Research in Environmental Sciences (CIRES), and the U.S. Department of Energy (DOE), and is facilitated by collaboration with the international Atmospheric Circulation Reconstructions over the Earth initiative. 20CR is the first ensemble of sub-daily global atmospheric conditions spanning over 100 years. This provides a best estimate of the weather at any given place and time as well as an estimate of its confidence and uncertainty. While extremely useful, version 2c of this dataset (20CRv2c) has several significant issues, including inaccurate estimates of confidence and a global sea level pressure bias in the mid-19th century. These and other issues can reduce its effectiveness for studies at many spatial and temporal scales. Therefore, the 20CR system underwent a series of developments to generate a significant new version of the reanalysis. The version 3 system (NOAA-CIRES-DOE 20CRv3) uses upgraded data assimilation methods including an adaptive inflation algorithm; has a newer, higher-resolution forecast model that specifies dry air mass; and assimilates a larger set of pressure observations. These changes have improved the ensemble-based estimates of confidence, removed spin-up effects in the precipitation fields, and diminished the sea-level pressure bias. Other improvements include more accurate representations of storm intensity, smaller errors, and large-scale reductions in model

[16]Fenner School of Environment & Society, Australian National University, Australia
[17]ARC Centre of Excellence for Climate Extremes, Australian National University, Australia
[18]National Centre for Atmospheric Science, Department of Meteorology, University of Reading, UK
[19]Direction de la Climatologie et des Service Climatiques, Météo-France, Toulouse, France
[20]Lamont-Doherty Earth Observatory of Columbia University, Palisades, New York, USA
[21]Hokkaido University, Sapporo, Japan
[22]States of Jersey Meteorological Department, Jersey, UK
[23]Hong Kong Observatory, Hong Kong, China
[24]National Institute of Water and Atmospheric Research, Auckland, New Zealand
[25]Department of Geography, Climatology, Climate Dynamics and Climate Change and Centre of International Development and Environmental Research, Justus Liebig University of Giessen, Germany
[26]Department of Environmental Science and Policy, Università degli Studi di Milano, Italy
[27]Department of Geography, University of South Carolina, USA
[28]Department of Physics, University of Toronto, Canada
[29]Department of Meteorology and Climatology, Nicolaus Copernicus University, Toruń, Poland
[30]Centre for Applied Climate Sciences, University of Southern Queensland, Australia
[31]Department of Oceanography, University of Cape Town, South Africa
[32]Centre for Interdisciplinary Studies of Montreal, McGill University, Montreal, Canada
[33]Deutscher Wetterdienst, Hamburg, Germany
[34]Instituto Dom Luiz, Faculdade de Ciências da Universidade de Lisboa, Portugal
[35]Environment and Climate Change Canada, Toronto, Canada
[36]NOAA Pacific Marine Environmental Laboratory, Seattle, Washington, USA
[37]Joint Institute for the Study of the Atmosphere and Ocean, University of Washington, USA

Correspondence
Laura C. Slivinski, NOAA/ESRL Physical Sciences Division, 325 Broadway, Boulder, CO 80305, USA.
Email: laura.slivinski@noaa.gov

bias. The 20CRv3 system is comprehensively reviewed, focusing on the aspects that have ameliorated issues in 20CRv2c. Despite the many improvements, some challenges remain, including a systematic bias in tropical precipitation and time-varying biases in southern high-latitude pressure fields.

KEYWORDS
data assimilation, reanalysis, surface pressure, 20CRv3

1 | INTRODUCTION

In order to study historical and contemporary weather events, including extremes, within a broader climate context, long time series of accurate, reliable, sub-daily atmospheric variables are essential. Retrospective analyses, or "reanalyses", take advantage of the benefits of past observations and modern weather forecast models by combining the two in a process called "data assimilation" (DA; Daley 1993). The idea of "reanalysis" arguably began in the early 19th century with Brandes' hand-drawn synoptic weather maps (Monmonier, 1999), and has matured significantly in the centuries since; Compo *et al.* (2006, 2011) give a detailed history. Historical reanalyses, which span a century or longer, act as a bridge between weather and climate since they are intended to capture individual weather events around the globe as well as larger climatic trends over many decades within the context of a single, consistent dataset (Slivinski, 2018).

In contrast to historical reanalyses, "modern" reanalyses generally only extend back to the 1950s, and more often only to 1979, when upper-air and satellite data became available for assimilation. These reanalyses include the European Centre for Medium-Range Weather Forecasts (ECMWF) interim reanalysis (ERA-Interim; Dee *et al.* 2011), the National Aeronautics and Space Administration (NASA) Modern-Era Retrospective analysis for Research and Applications version 2 (MERRA-2; Gelaro *et al.* 2017), the 55-year Japanese Reanalysis (JRA-55; Kobayashi *et al.* 2015), and the reanalysis produced jointly by the US National Centers for Environmental Prediction (NCEP) and the National Center for Atmospheric Research (NCAR), the NCEP-NCAR Reanalysis (Kalnay *et al.*, 1996; Kistler *et al.*, 2001), among others. (Fujiwara *et al.* 2017 give a review of reanalysis systems.) At present, long-term studies using modern reanalyses are restricted to span as few as 40–60 years, preventing in-depth investigation of infrequent extreme weather and climate events. Another difficulty is that significant changes to the observing system, such as the introduction of satellite data, can yield non-climatic discontinuities in some reanalysis fields, including an apparent shift in the tropical divergent circulation (Kinter *et al.*, 2004) and trends in temperature, integrated water vapour, kinetic energy, and precipitation (Bengtsson *et al.*, 2004; Bosilovich *et al.*, 2011; Zhang *et al.*, 2012). In order to avoid such artifacts, historical reanalyses that span at least a century assimilate only near-surface conventional observations, which have been available for the entire time period: specifically, these include surface pressure and marine winds.

The NOAA-CIRES Twentieth Century Reanalysis (20CR) marked the introduction of recent efforts to generate historical reanalyses, as it was the first reanalysis to assimilate only surface pressure observations (Compo *et al.*, 2011). Since then, the range of studies to use these types of data has grown, and other centennial reanalyses were developed that assimilated these data. ECMWF produced ERA-20C (Poli *et al.*, 2016), an atmospheric reanalysis spanning 1900 to 2010 that assimilated surface pressure as well as marine winds, and CERA-20C (Laloyaux *et al.*, 2018), which utilizes a coupled ocean–atmosphere model and spans 1901 to 2010. In addition, NOAA and CIRES produced an update to the 20CR version 2 described by Compo *et al.* (2011) that spanned 1871 to 2012; this update, 20CR version 2c (20CRv2c; Giese *et al.*, 2016 and detailed below), extended back to 1851 and ameliorates several issues with 20CRv2. Finally, the latest 20CR version 3 (20CRv3) is currently being produced by NOAA, CIRES, and DOE. It is expected to extend back to 1836 and be released in 2019.

Historical reanalyses have broad areas of application because they span time-scales of weather to climate by providing sub-daily estimates of the Earth system with global coverage for a century or longer. These datasets have been utilized in studies including: climate change (e.g. Compo *et al.*, 2013; Huang *et al.*, 2016); climate dynamics (e.g. Huang *et al.*, 2017); trends in hurricanes (e.g. Burn and Palmer, 2015), extratropical cyclones (e.g. Wang *et al.*, 2013, 2016), and extremes in temperature and precipitation (e.g. Donat *et al.*, 2016); blocking (e.g. Häkkinen *et al.*, 2011; Rohrer *et al.*, 2018); individual case-studies of particular storms (e.g. Moore and Babij, 2017); historic climatology in remote regions (e.g. Lorrey and Chappell, 2016); El Niño (e.g. Giese *et al.*, 2010; Deser *et al.*, 2017); the Madden–Julian Oscillation (e.g. Klotzbach *et al.*, 2016); convergence zone activity (e.g. Lorrey *et al.*, 2012; Harvey *et al.*, 2019); seasonal and climatic responses to volcanic eruptions (e.g. Brohan *et al.*, 2016; Paik and Min, 2017); weather typing (e.g. Jones *et al.*, 2013, 2016); and the emerging field of decadal climate prediction (e.g. Mueller *et al.*, 2014), among many others.

A key aspect for informed application of reanalyses is properly accounting for their uncertainty (e.g. Parker, 2016). Comparing different reanalyses that span similar time periods is one way to cross-validate the datasets and determine a "meta-confidence" by agreement or disagreement among the datasets. It is also important that each historical reanalysis dataset is as accurate as possible, both in terms of past climate state estimates as well as internal quantification of its uncertainty (as measured by ensemble standard deviation or "spread", for instance). This internal quantification of uncertainty is used by the DA system during the production of the reanalysis to make the best use of the observations and prior background information, but is also important to the end-users of the reanalysis. As an example, a historical reanalysis may display a long-term trend in one variable that, according to the quantified uncertainty of the dataset, is significant. However, researchers may be unaware that the trend is an artificial one due to a bias in the observations, and appears to be significant solely due to errors in the uncertainty estimate. Continuing to work towards more reliable historical reanalyses allows

studies on all time-scales, such as those listed above, to avoid erroneous conclusions and make use of the best data possible.

In this vein, despite several major improvements from 20CRv2 to 20CRv2c, certain issues remain. While some are obvious, such as artificial large-scale trends and a lack of certain major storm systems, others are more subtle, such as suboptimal usage of observations and inaccurate estimates of confidence. These problems can hinder the effectiveness of 20CRv2c for climate analysis applications. Investigations into many of these issues occurred prior or in parallel to development of version 3, informing the implementation of particular algorithms that are expected to improve the efficacy of the reanalysis. In other cases, version 3 will likely benefit from general improvements and upgrades to the system, as well as a larger observational database. This work discusses how the significant issues in version 2c are addressed, as well as other upgrades to the version 3 system. Preliminary results with the 20CRv3 dataset shown here will focus on several test periods between 1851 and 2002, and are intended to be representative of different time periods (in terms of quality, confidence, observational network density, biases, etc.) Results from the complete 20CRv3 dataset and deeper investigations of it on climatic and synoptic scales are left for future work. Unless otherwise noted, all maps shown below are plotted at the native resolution of the dataset. Finally, we emphasize that many updates to the 20CRv3 system were made simultaneously, so a single improvement in this preliminary 20CRv3 data can rarely be attributed to a specific change in the system.

The Twentieth Century Reanalysis system is described in detail in Section 2. Aspects of the system that changed from 20CRv2c to 20CRv3 are highlighted, as well as features of the version of the NCEP Global Forecast System (GFS) coupled atmosphere–land model used. Section 3 discusses several large-scale issues in the confidence derived from ensemble spread and in the biases of sea-level pressure (SLP), precipitation, and wind in 20CRv2c. Preliminary results suggest that updates to the forecast model and DA algorithm will improve the confidence estimation and reduce most of these biases in 20CRv3. In addition to addressing known issues, other developments in the version 3 system are expected to result in further improvements. As shown in Section 4, updates to the localization procedure, quality control (QC), and observation errors will likely improve the use of observations and result in more accurate representations of variability and extremes, such as tropical cyclones. Despite significant improvements across the board, several issues remain in the 20CRv3 system. These are discussed in Section 5. Section 6 concludes with a discussion and final remarks.

2 | SYSTEM OVERVIEW

In several basic ways, each iteration of the Twentieth Century Reanalysis system remains the same as that proposed originally by Compo *et al.* (2006). First, modern weather forecast models are used to generate the atmospheric background fields given prescribed sea surface temperature (SST) and sea ice concentration fields. Second, an ensemble method assimilates historical observations to update the background fields, yielding analysis fields. Ensemble methods are particularly useful as they allow for estimates of uncertainty and confidence via ensemble spread (e.g. ensemble standard deviation) as well as an estimate of the atmospheric state via the ensemble mean. Finally, surface pressure values are the only type of observations that are ever assimilated. Sea ice and SST observations are implicitly included in the reanalysis in the form of boundary conditions and can guide the model to represent large-scale climate features, but they are not assimilated. While the temporal frequency, spatial density, and quality of the surface pressure observations have changed over time as a result of developments in instrumentation and theory (Middleton, 1964), the 20CR system assumes that the most important part of the observation error is its so-called "error of representativeness" (Lorenc, 1986; Janjić and Cohn, 2006). Observation errors are therefore assumed to be constant in time; Section 4.2 gives more discussion. Feedback from reanalysis datasets that assimilate these observations can be used to improve this estimate in the future (e.g. Poli *et al.*, 2015; Laloyaux *et al.*, 2018).

For purposes of comparison, the 20CRv2c system outlined by Giese *et al.* (2016) is detailed in Appendix A, and the 20CRv3 system is detailed here. To address significant issues in the 20CRv2c dataset, and as a result of general progress in the fields of modelling and DA, several aspects of the 20CR system were updated before producing 20CRv3. Broadly, 20CRv3 will benefit from an improved, higher-resolution model, a larger observational database, updated DA methods, and a larger ensemble size. The atmospheric model used in 20CRv3 has been updated to the 2017 version of the NCEP GFS with a resolution of total spherical wavenumber 254 (about 0.5° horizontal resolution) and 64 vertical hybrid sigma-pressure levels; differences between the version of the GFS operational in autumn 2017 and the version used for 20CRv3 are detailed in Appendix A. Additionally, the version 2c system allowed the assimilation to update the dry air pressure, resulting in a feedback loop with biased observations that caused significant artificial trends in the mid-19th century (Section 3). In version 3, the dry air pressure was held fixed in the forecast and analysis steps.

The 20CRv2c dataset began in 1851 due to the availability of its prescribed sea ice fields. The addition of more 19th century observations available to the 20CRv3 assimilation system, as well as early investigations of confidence and forecast errors (not shown), suggested that 20CRv3 could span further back in time than 20CRv2c, given appropriate boundary conditions. 1804 is the first year that every 6-hr window has at least one observation (globally) to be assimilated. Due

to computational and storage resource limitations, 1836 was the earliest year that 20CRv3 could be produced. Experiments for the years 1804–1835 are ongoing.

The 20CRv3 dataset is expected to consist of two overlapping sub-versions: 20CRv3si (1836–2012) and 20CRv3mo (1981–2015), where the only difference between the two sub-versions is the prescribed SSTs. 20CRv3mo prescribes SSTs from HadISST2.2 (Rayner *et al.*, 2006; Poli *et al.*, 2016; Laloyaux *et al.*, 2018), which consists of an ensemble of 5-day average SST fields interpolated to daily resolution. This interpolation is cubic, with the coefficients adapted to the autocorrelation of the data (Kwon *et al.*, 2004). Of the ten members available, two of the ensemble members had quite different bias adjustments from the others; thus, 20CRv3mo only uses the remaining eight members as boundary conditions. 20CRv3si prescribes SSTs from the pentad, linearly interpolated to daily, eight-member Simple Ocean Data Assimilation with sparse input version 3 (SODAsi.3) ensemble that itself used 20CRv2c fields as atmospheric boundary conditions and forcing (Giese *et al.*, 2016). The SODAsi.3 SSTs used for 20CRv3si were seasonally adjusted to the 1981–2010 HadISST2.2 daily climatology. For both versions, each of the eight distinct SST ensemble members was duplicated ten times to create a total of 80 members. Thus, the first, 9th, 17th, 25th, ..., and 73rd members of the 20CRv3 ensemble have the same SST forcing, and the second, 10th, 18th, 26th, ..., and 74th members have the same SST forcing as each other (but different from the first set), and so on. Sea ice concentrations were specified from HadISST2.3, which is identical to HadISST2.2 (Titchner and Rayner, 2014) from 1972 onwards. From 1850 to 1971, HadISST2.3 specifies Arctic sea ice extent from the Sea Ice Back To 1850 dataset (SIBT1850; Walsh *et al.*, 2016). Prior to 1850, sea ice extent and concentration are specified as the 1860–1891 HadISST2.3 climatology.

Thanks to international efforts facilitated by the Atmospheric Circulation Reconstructions over the Earth (ACRE) initiative (Allan *et al.*, 2011) and many volunteer efforts, there are millions more observations assimilated in 20CRv3 than in 20CRv2c. This represents an average of 5% more available observations per assimilation cycle in recent periods (after about 1930), and up to 25% more available observations per cycle in earlier years. The new observational dataset, the International Surface Pressure Databank (ISPD) version 4.7 (Compo *et al.*, 2015; Cram *et al.*, 2015), blends surface pressure and sea-level pressure (SLP) from the Integrated Surface Database (ISD; Lott *et al.*, 2008; Smith *et al.*, 2011) with additional station observations, archived and previously undigitized terrestrial data submitted to the ISPD from international ACRE partners, pressure reports for tropical cyclones from version V03r10 of the International Best Track Archive for Climate Stewardship (IBTrACS; Knapp *et al.*, 2010; Kruk *et al.*, 2010) combined with additional Pacific tropical cyclone data (Kubota, 2012), and marine observations from the International Comprehensive Ocean–Atmosphere Data Set (Worley *et al.*, 2005; Woodruff *et al.*, 2011; Freeman *et al.*, 2017) ICOADS3+ version 2. The latter is our own improvement to ICOADS3 that includes recently digitized and better positioned and quality-controlled observations from ACRE-recovered expeditions, OldWeather.org, and the Australian Weather Detective project (Spencer *et al.* 2019; https://github.com/oldweather/ICOADS3.plus/releases, accessed 9 July 2019, and Appendices A, B).

Unlike 20CRv2c, which used a 56-member ensemble Kalman filter with a digital filter applied to the background forecast, 20CRv3 assimilates observations with an 80-member ensemble Kalman filter that utilizes a four-dimensional incremental analysis update (Bloom *et al.*, 1996; Lei and Whitaker, 2016) and no digital filtering (Section 3 and Appendix A). Additionally, 20CRv2c interpolated station pressure observations to the model surface prior to assimilation, while 20CRv3 uses the more typical procedure and assimilates them at the observation level, absorbing the vertical interpolation of the background forecast into the observation operator (**H**). As will be discussed in Section 4, 20CRv3 includes a nonlinear QC algorithm for the observations, an adaptive localization algorithm, an inflation method based on relaxation-to-prior-spread, and an offline bias correction for marine observations prior to 1871 (Appendices B–D give more details). 20CRv3 also includes an updated bias correction for station data over land: these biases are "learned" over a 60-day time period. That is, they are calculated as the average difference between the observation and the first guess over the 60-day window (with a minimum of 31 days' worth of data in the window) prior to the current assimilation step; if significant, these differences are subsequently removed from the observation at the step prior to assimilation (Compo *et al.*, 2011, their appendix B, give more details). Finally, the baseline observation errors used in 20CRv3 are given in Table 1. Row 4 ("station") refers to observations of surface pressure, while row 5 ("SLP only") refers to stations that only reported pressure reduced

TABLE 1 Platform-dependent baseline observation errors (hPa) used in the 20CRv3 system. Note that only surface pressure data are assimilated, including from radiosonde and dropsonde observing platforms. "SLP only" refers to stations that do not report surface pressure, only sea-level pressure

Type	Error (hPa)
Radiosonde	1.2
Dropsonde	2.0
Marine	2.0
Station	1.2
Station (SLP only)	1.6
Tropical cyclones	2.5

SLP = sea-level pressure.

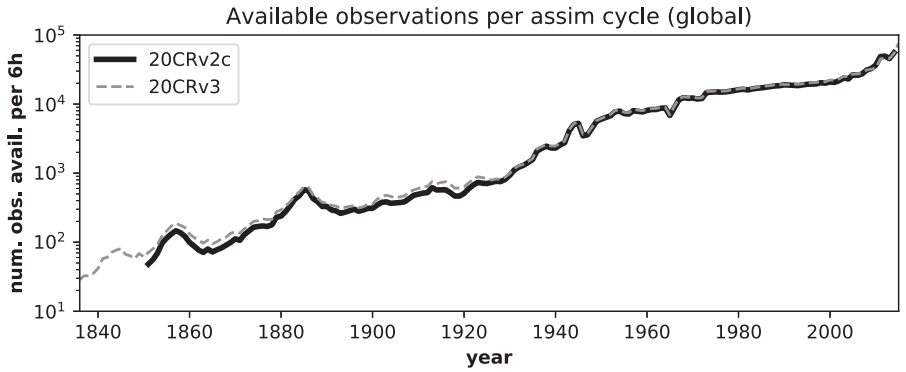

FIGURE 1 Time series of the annual average number of observations available to be assimilated globally per 6-hr window within 20CRv2c (ISPDv3.2.9, solid black) and 20CRv3 (ISPDv4.7, dashed grey)

to sea level. Observation errors are increased by 0.001 hPa for each one-meter difference between the observation elevation and the model orography. These are the same errors used in 20CRv2c, with the exception of tropical cyclone data (Table A1 in Appendix).

3 | ADDRESSING ISSUES IN 20CRV2C

The 20CRv3 dataset will build on two previous efforts: 20CRv2 and 20CRv2c. The 20CRv2 dataset represented an important step forward for weather and climate research because it filled a need for a consistent, long-term, sub-daily gridded atmospheric dataset using instrumental observations. As of the time of writing, the paper describing the 20CRv2 dataset (Compo *et al.*, 2011) has more than 2000 citations (Google Scholar; accessed 4 February 2019). While useful, the 20CRv2 dataset has several issues, including a mis-specification of polar sea ice that resulted in positive near-surface temperature biases (Brönnimann *et al.*, 2012) and inhomogeneities associated with variations in observation density and its covariance inflation algorithm prior to 1952 (Ferguson and Villarini, 2012).

The 20CRv2c dataset was an effort to address those issues, use a novel SST specification, and include additional observations compared to 20CRv2. However, as more studies delved into different aspects of 20CRv2c, its limitations became apparent. Simultaneously, the many studies using 20CRv2c motivated further data rescue efforts, and the number of pressure observations available to be assimilated grew significantly, particularly in early years. Figure 1 illustrates the global annual average number of observations available to be assimilated in a 6-hr window of 20CRv2c (solid black) and 20CRv3 (dashed grey). Here, "available" refers to observations that were rescued, digitized, externally quality controlled, and blended into the version of the ISPD used in the given reanalysis; it includes observations that may be flagged or thinned by the internal 20CR QC system (Section 4 and Appendix C give details.)

A new version of the 20CR system could make use of this growing set of observations, as well as general progress in modelling and DA methods, and would provide a significantly improved dataset. Major issues in 20CRv2c, including inaccurate representations of uncertainty as well as large-scale biases and artifacts in SLP, precipitation, and wind, also informed and motivated the development of the 20CRv3 system.

3.1 | Estimation of confidence

In order to make conclusions about the significance of trends, signals, and extrema from reanalyses, we must be able to quantitatively measure confidence in the datasets. A defining characteristic of 20CR is its use of an ensemble DA method, which yields both a single best estimate of the analysis (the mean) as well as a quantification of the uncertainty around that estimate via the ensemble spread (the standard deviation). More spread implies more uncertainty, and less spread implies less uncertainty. In general, the uncertainty in the ensemble mean as an estimate will correlate negatively with the density of the available observational network. In addition, the variability of the ensemble mean in time, or temporal spread (Equation 1), can be used as an estimate of climatological uncertainty. At time t_k, the temporal spread of the ensemble mean over a window of length ΔT is given by:

$$\text{temp. spread}(t_k) = \left[\frac{1}{N_{\text{time}}-1}\sum_{t=t_k-\Delta T/2}^{t_k+\Delta T/2}\left(x_{\text{ensmean}}(t)-\overline{x_{\text{ensmean}}}\right)^2\right]^{1/2}, \quad (1)$$

where N_{time} is the number of time steps in the window $[t_k - \Delta T/2, t_k + \Delta T/2]$, $x_{\text{ensmean}}(t)$ is the area-averaged ensemble mean of the variable of interest (for example, SLP) at time t, and $\overline{x_{\text{ensmean}}}$ is the time-average of the ensemble mean $x_{\text{ensmean}}(t)$ over the time window.

As an example, Figure 2 includes time series of uncertainty in SLP over the zonal band from 65 to 40°S calculated

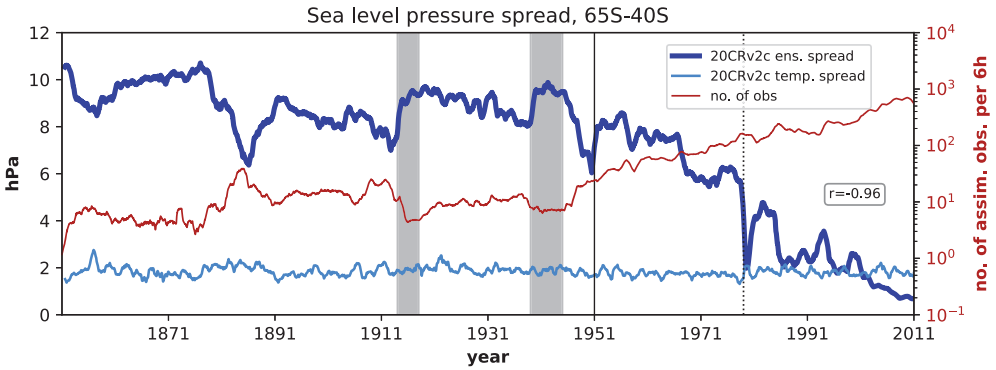

FIGURE 2 Time series of ensemble spread (bold dark blue curve) and temporal spread of the ensemble mean (thin light blue curve) for sea-level pressure from 20CRv2c averaged over the zonal band from 65 to 40°S. Number of observations assimilated per 6-hr window in this region is shown in red (right-hand axis). A 1-year running average was applied to all curves. Correlation r is calculated between the smoothed ensemble spread and the smoothed logarithm of the number of assimilated observations. Notable years are emphasized with vertical lines and shading; see text for details

from 20CRv2c. This region is particularly important for investigations of anthropogenic CO_2 uptake, and long time series are needed for studies of its decadal variability (as discussed in e.g. Landschützer *et al.*, 2015). The analyzed SLP ensemble spread is plotted (thick dark blue curve) along with the temporal spread of the analyzed SLP ensemble mean (thin light blue curve) and the number of observations assimilated per 6-hr window (thin red curve, right-hand axis) in this region. The temporal spread is calculated as the standard deviation of the ensemble mean across a centred time window using Equation (1) with $\Delta T = 61$ days, and all time series have a 1-year running average applied. This region has relatively few observations available (compare with Figure 1), and the effects of World War I (1914–1918) and World War II (1939–1945) are particularly striking (shaded grey regions), as is the First GARP (Global Atmospheric Research Program) Global Experiment in 1979 (dashed line). The correlation r between the ensemble spread and the log of the number of observations assimilated per window is –0.96, demonstrating the strong inverse relationship between the ensemble spread and the observational network density.

However, ensemble spread is only an estimate of uncertainty, and it is not always reliable. For instance, a well-known issue with the EnKF is the tendency for ensembles to "over-tighten" towards the mean, resulting in an ensemble spread that is overconfident and ultimately in filter divergence (that is, when the background ensemble standard deviation approaches 0 and the ensemble is unable to use information about observations; Anderson and Anderson, 1999; Whitaker and Hamill, 2002). A common method to address this problem is "covariance inflation"; generally, this refers to artificially increasing the ensemble spread by, for example, applying a multiplicative factor greater than 1 to the ensemble covariance. When many observations are assimilated, the ensemble is more prone to collapse, and thus requires more inflation. In 20CRv2c, a simple multiplicative inflation factor

TABLE 2 Covariance inflation parameters used in 20CRv2c as a function of latitude and year. A value of 1 corresponds to no inflation

Years	NH	Tropics	SH
1851 – 1870	1.01	1.01	1.01
1871 – 1890	1.05	1.01	1.01
1891 – 1920	1.09	1.02	1.01
1921 – 1950	1.12	1.03	1.02
1951 – 2014	1.12	1.07	1.07

NH = 90–30°N; Tropics = 30°N–30°S; SH = 30–90°S.

(Anderson and Anderson, 1999) was applied to the ensemble covariance matrix at each step; this factor was predefined based on year and latitude. Table 2 shows the inflation parameters used in 20CRv2c. These time periods were chosen to loosely reflect availability and density of observations: for example, there were few National Meteorological Services organized prior to 1870, and thus the observational network was relatively sparse. The period 1871–1890 represents a transition period; with such developments as the founding of the International Meteorological Committee in 1873, the network of observations in the Northern Hemisphere becomes denser. Conversely, the Southern Hemisphere observation network remains relatively sparse into the 20th century.

While this method ensured that larger inflation parameters were applied when the observation network was more dense (e.g. in the Northern Hemisphere and in modern time periods), the abrupt changes in the parameters are responsible for artificial signals in the time series of uncertainty. The spike in ensemble spread in 1951 (solid black line in Figure 2) is an artifact of the multiplicative inflation algorithm used in 20CRv2c; this is the year the inflation parameter in the Southern Hemisphere increased from 1.02 to 1.07 (Table 2), and there is no corresponding decrease in number of observations assimilated. In fact, 1951 marks an increase in assimilated SLP observations; this originally motivated increasing the inflation parameter in that particular year.

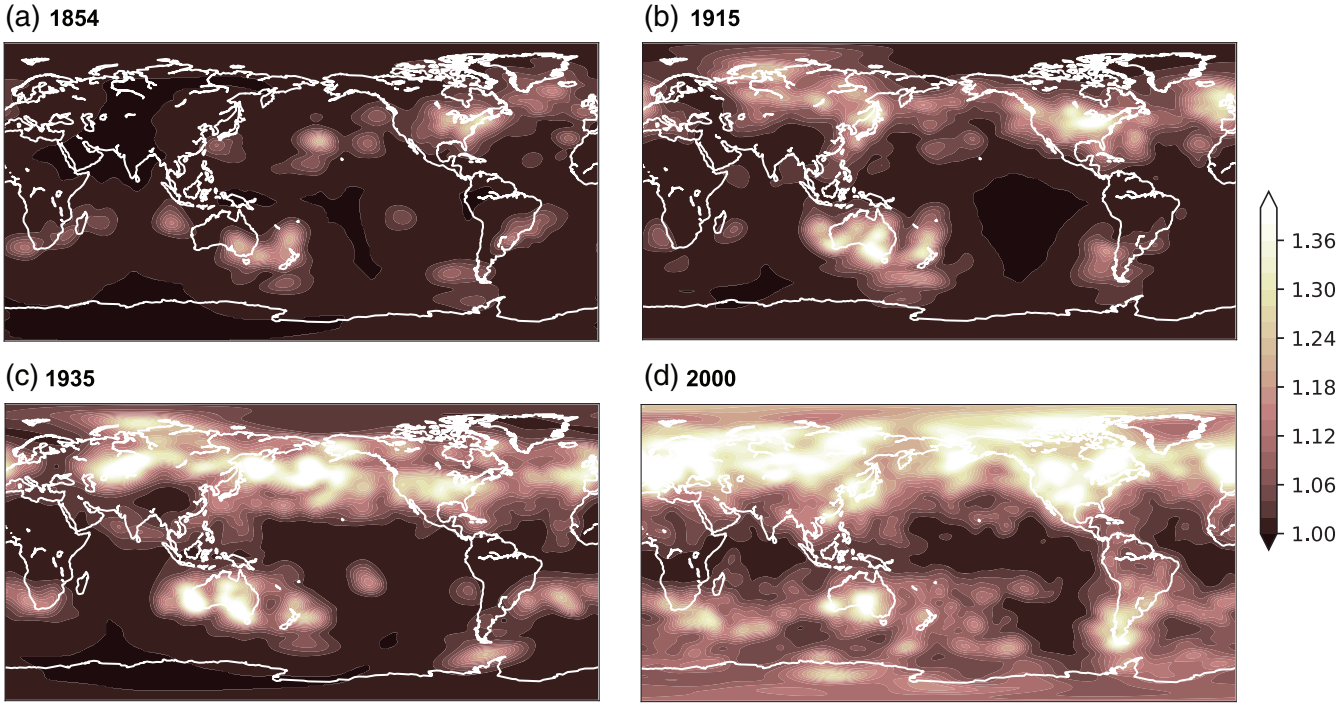

FIGURE 3 Maps of the adaptive inflation parameter λ_{inf} (unitless) used in the 20CRv3 system for 0000 UTC on 1 September in (a) 1854, (b) 1915, (c) 1935, and (d) 2000. A value of 1 corresponds to no inflation. Note the relationship of λ_{inf} with the observation network density

Another issue demonstrated by Figure 2 is the under-confidence of the ensemble spread. As discussed earlier, the temporal spread can be used as a proxy for a climatological spread. Until the 1980s, the ensemble spread is larger than the temporal spread, suggesting that the ensemble was less confident than a climatological estimate. The inverse also occurs: the fixed inflation algorithm can result in too little inflation over data-rich regions, leading to overconfidence in these areas (not shown).

The version 3 system uses an improved inflation algorithm referred to as relaxation-to-prior-spread (Whitaker and Hamill, 2012). Using this algorithm, the inflation adapts to the observation network density. When there are few observations, the ensemble spread is hardly changed; when there are dense observations, the ensemble spread is "relaxed" back to the prior spread, by an amount λ_{inf}. For every model grid point (x, y) and analysis time t, the inflation parameter λ_{inf} is given by:

$$\lambda_{\text{inf}}(x,y,t) = p_{\text{relax}}\left(\frac{\sigma_{\text{b}}(x,y,t) - \sigma_{\text{a}}(x,y,t)}{\sigma_{\text{a}}(x,y,t)}\right) + 1, \qquad (2)$$

where $\sigma_{\text{b}}(x,y,t)$ is the standard deviation of the background ensemble, $\sigma_{\text{a}}(x,y,t)$ is the standard deviation of the analysis ensemble before inflation, and p_{relax} is a relaxation parameter that can vary from 0 (no inflation) to 1 (inflate to prior spread). The ratio of ensemble spread in Equation (2) implicitly depends on the density of the observation network in that region: a dense network will result in a smaller analysis ensemble spread, and thus a larger inflation parameter, and vice versa. Initial tests with the 20CRv3 system used $p_{\text{relax}} = 0.9$ globally. These tests (not shown) suggested that this was too large in the Southern Hemisphere, as the uncertainty was larger than a climatological uncertainty. Thus, the final 20CRv3 system uses $p_{\text{relax}} = 0.9$ for 20°S–90°N and $p_{\text{relax}} = 0.7$ for 90–30°S. In the transition zone 30–20°S, p_{relax} varies linearly from 0.7 to 0.9. These values of p_{relax} do not change in time. Figure 3 shows representative examples of the adaptive inflation parameter λ_{inf} from four different years; a value of 1 is equivalent to no inflation.

The GFS model in 20CRv3 uses stochastic physics (Appendix A), which also contributes to the ensemble spread. This effect is particularly strong in the Tropics, which reduces the need for inflation in this region (e.g. Figure 3d). Outside the Tropics, the inflation factor depends on the observation network density: over the US and Europe, and throughout the Northern Hemisphere in recent years, the inflation factor is larger than elsewhere. Note also that the range of inflation parameter values used in 20CRv3 is much larger than was prescribed in 20CRv2c (compare Figure 3 and Table 2).

Figure 4 illustrates the result of these changes in terms of the "confidence" in fields of SLP from versions 2c and 3 of 20CR during selected early 20th century boreal winters. Here, "confidence" is defined as the difference of the normalized time-averaged ensemble standard deviation from 1:

$$conf = 1 - spread_{\text{ens}}/spread_{\text{clim}}, \qquad (3)$$

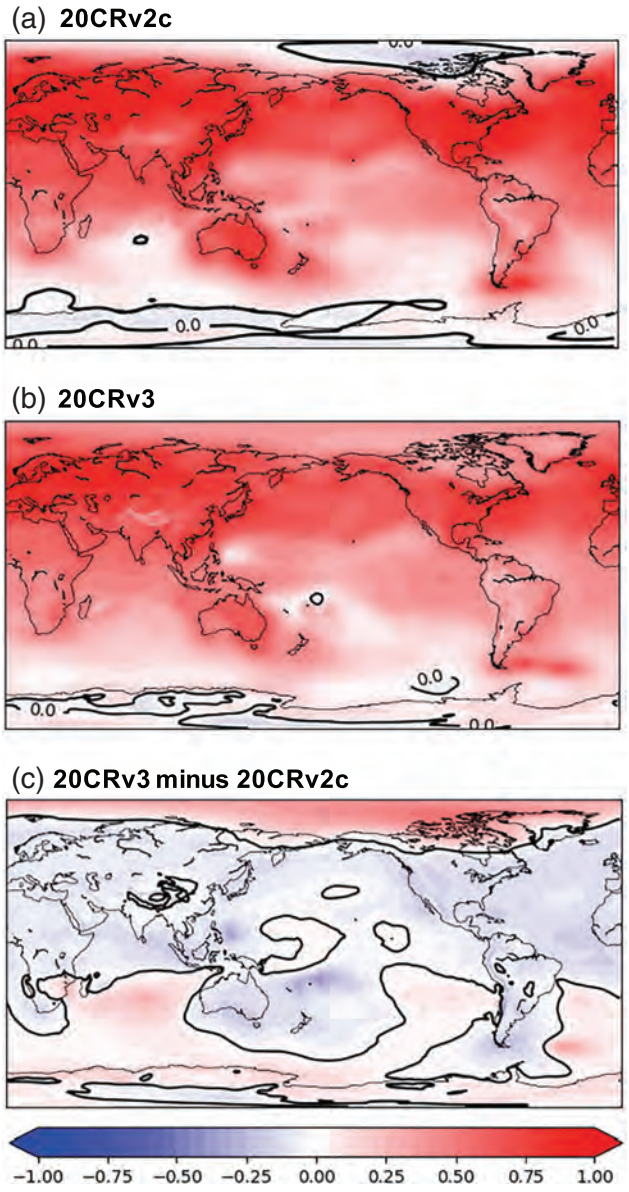

FIGURE 4 Maps of normalized confidence of SLP averaged over Jan-Feb-Mar for 1916–1918 from (a) 20CRv2c and (b) 20CRv3, and (c) the difference (20CRv3 minus 20CRv2c). In (a, b), zero (black contour) represents climatological uncertainty, blue represents less certainty than climatology, and red represents more certainty. In (c), red represents an increase in confidence from 20CRv2c to 20CRv3, and blue a decrease

where $spread_{\mathrm{ens}}$ is the time-averaged standard deviation of the ensemble of analyzed SLP from the stated version of 20CR, and $spread_{\mathrm{clim}}$ is the temporal standard deviation of the 20CRv2c ensemble mean 6-hr SLP over Jan-Feb-Mar from 1981–2010. In other words, $spread_{\mathrm{clim}}$ represents an estimate of the inherent weather variability; it is assumed to be time-invariant and independent of ensemble spread. Thus, a confidence value of zero (denoted by black contours in Figure 4a, b) denotes an ensemble spread identical to the climatological spread; greater confidence implies more certainty than climatology, and negative confidence implies less certainty than climatology. Aside from interannual variations in weather variability (e.g. Compo *et al.* 2001), the minimum confidence value would be zero.

These maps demonstrate features of the new estimates of confidence in 20CRv3. In particular, there is more certainty over the high Arctic latitudes in version 3 (red shading in Figure 4c in this region) than simply using a climatological mean as the analysis. There is also more spread over the densely observed regions of North America and Europe (blue shading in Figure 4c in these regions). Results (not shown) using independent observations from UK Daily Weather Reports comparing expected and actual errors suggest that this is an improvement, as the 20CRv2c analyses are overconfident over Europe in the 1900s. Similar overconfidence is found for 20CRv2c first-guess fields (not shown). We expect that the results would be similar for independent observations over North America. In contrast, the larger-than-climatological uncertainty over the high southern latitudes has been reduced but not eliminated, despite the decrease in p_{relax} discussed above. There is also a decrease in confidence in 20CRv3 throughout much of the Tropics; this may be due to the stochastic physics described above. The decrease in confidence throughout the midlatitude oceans suggest that further experiments with p_{relax} may be necessary in the future. While many of the differences are likely due to the new adaptive inflation algorithm, recall that 20CRv3 uses an 80-member ensemble, as opposed to the 56 members in 20CRv2c. The larger ensemble, as well as other updates to the 20CRv3 system, may also have contributed to greater consistency between the quantified confidence of version 3 and prior expectations.

3.2 | Global sea-level pressure bias

Another significant issue in 20CRv2c, an SLP bias prior to the 1870s, prevented this dataset from being as useful as it could have been for its full span. This bias is evident in globally averaged time series of SLP (Figure 5, blue curve) for most years prior to 1870. Relative to several reanalyses of varying timespans, including ERA-Interim (orange), the historical reanalyses ERA-20C (green) and CERA-20C (gold), and a 56-member ensemble of simulations with the same version of the GFS used in 20CRv2c but without assimilation ("no DA"; red), the global SLP from 20CRv2c is as much as 2–4 hPa too low during the period of concern. Shading on Figure 5 represents one standard deviation when ensemble estimates are available; note that the 20CRv2c spread in the biased period is still several hPa away from the "no DA" mean and standard deviation.

The cause is revealed to be biased ship observations in the mid-19th century, first reported by Todd Mitchell at a marine data workshop (Diaz *et al.*, 2002; Allan and Ansell, 2006; Ansell *et al.*, 2006), combined with the 20CRv2c system

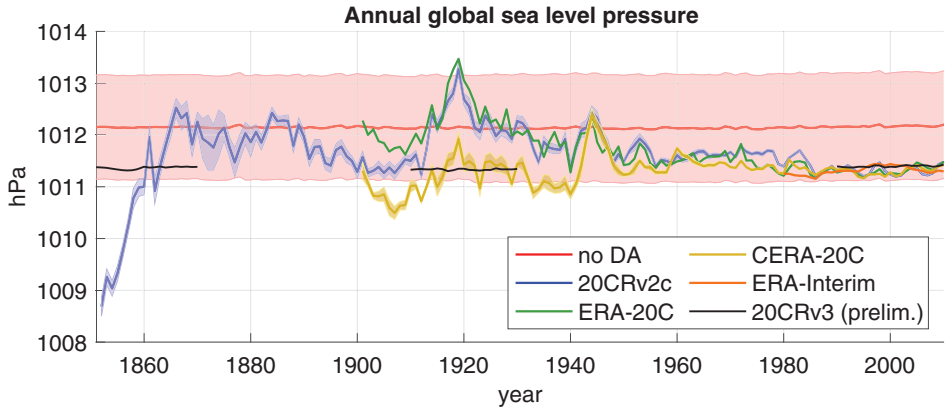

FIGURE 5 Time series of annually averaged global SLP from several reanalyses, as well as a non-assimilating (no DA) ensemble of model simulations

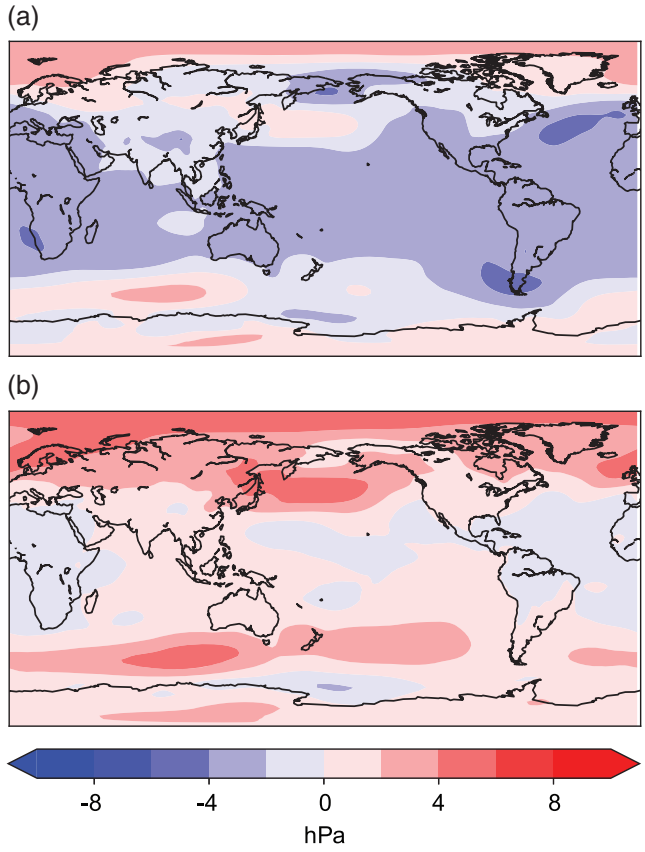

FIGURE 6 Maps of analyzed sea-level pressure anomalies (with respect to the analyzed 20CRv2c 1981–2010 climatology) for 1851–1853 of (a) 20CRv2c and (b) an identical experiment with the 20CRv2c system which assimilated 10% fewer ship observations

allowing the global dry air pressure to be updated during the assimilation cycle. The 20CRv2c system assimilated many low-biased ship observations throughout the world oceans. During the assimilation, the global SLP field and the global dry air pressure were updated to be lower and more consistent with these biased observations. This resulted in a feedback cycle, as the biased observations continued to be assimilated in the mid-19th century and the global SLP and dry air pressure fields continued to be lowered. While 20CRv2c included a bias correction to land stations, it did not include a marine observation bias correction algorithm. Figure 6a shows a map of the 1851–1853 time-averaged SLP anomaly fields from 20CRv2c analyses: note the widespread negative anomalies, particularly over the oceans. Figure 6b differs from Figure 6a in that it shows an experiment that assimilated about 10% fewer ship observations than 20CRv2c, but with an otherwise identical set-up. Overall, the anomalies are less negative, demonstrating that assimilating more ship observations from 1851–1853 negatively biased the globally averaged analyzed SLP by as much as 4 hPa. Note that these ships were not chosen on an *a priori* basis: this illustrates the strong effect of a small number of biased observing platforms within a cycled DA system.

Two improvements in 20CRv3 will address this issue. First, the global dry air pressure can no longer be updated within the assimilation: instead, it is specified at 98.3050 kPa (Trenberth and Smith, 2005). This prevents the feedback loop with the biased ship observations that allowed the global SLP bias to persist for nearly two decades of 20CRv2c data. Second, to directly address the observation bias, a correction is applied to marine observations prior to 1870. Investigations into the individual observations found that the negative bias is not consistent across different voyages in this time period, suggesting that a single bias correction for all marine observations in this time period would not be sufficient. Thus, a bias for each individual ship is calculated as the mean deviation from the 20CRv2c 1981–2010 climatology, and subtracted from the marine observations prior to assimilation (Appendix B gives more details).

Figure 7 illustrates a test of these new procedures. Figure 7a shows the 20CRv2c SLP annual anomaly for 1854; note the consistently negative differences throughout the Tropics and midlatitudes. The effect of constraining dry air pressure in the version 3 system without bias correcting the observations is shown in Figure 7b. There are still negative

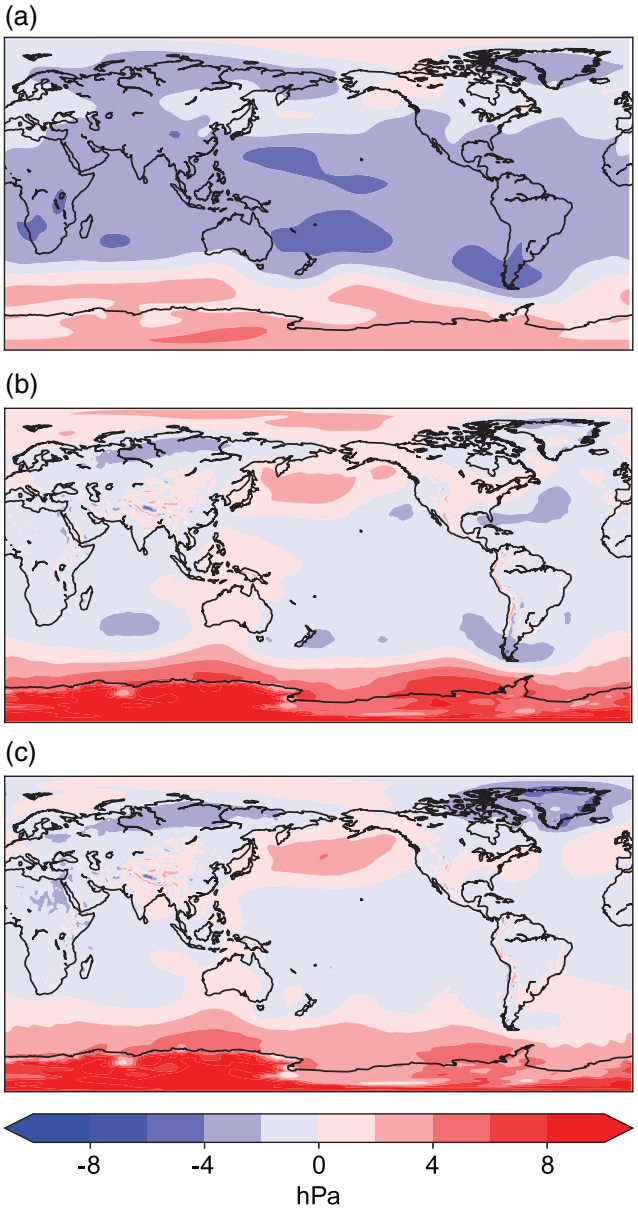

FIGURE 7 Maps of the year 1854 annually averaged sea-level pressure anomaly (with respect to the 1981–2010 climatology) of (a) 20CRv2c (without dry air mass specification), (b) preliminary version 3 tests without ship bias correction (with dry air mass specification), and (c) preliminary version 3 tests with ship bias correction (with dry air mass specification)

anomalies in the highly trafficked regions of the ocean (around Cape Horn, South America; the North Pacific Ocean off the coast of the US; and the North Atlantic Ocean). In order to retain the fixed dry pressure, this leads to an increased SLP anomaly where there are fewer observations, particularly around the poles. Figure 7c includes both the fixed dry pressure and the bias-corrected ship observations. The negative anomalies in high-density marine regions are now almost entirely removed, and while the positive anomaly over the high southern latitudes remains, it has been diminished. The black curves in Figure 5 represent the annual global SLP from 20CRv3 during three test periods, and demonstrate the large-scale effects of these changes to the 20CRv3 system.

As with many bias correction schemes, it is possible that this method is removing real signals from historical observations by forcing them towards a modern climatology. For example, the negative SLP anomalies in the southern midlatitudes prior to bias correction (Figure 7b) are assumed to be effects of biased observations from ships, as these anomalies are strongest in heavily trafficked shipping lanes and whaling areas. However, this pattern could be a real climatological shift in wave number 3 of the zonal flow in the Southern Hemisphere (e.g. van Loon and Jenne, 1972; Raphael, 2004) and would be erroneously removed by the bias correction scheme (Figure 7c). Nevertheless, in the absence of more information about these pressure observations or independent reconstructions of the mid-19th century SLP fields for validation, this procedure provides an improvement over uncorrected marine observations leading to spurious SLP trends (e.g. Figure 5). Deeper investigations into the cause of this observational bias (such as changes in meteorological logs or barometer-correction practices over the period 1850–1860) could allow for more realistic bias correction schemes in the future.

3.3 | Artifacts in precipitation and wind

While the global SLP trend prior to 1870 in Figure 5 could be attributed in some way to the observations, other artifacts can be traced back to the assimilation method. One example stems from the use of a digital filter (Lynch and Huang, 1992; Huang and Lynch, 1993) in the forecast step of 20CRv2c which was implemented to temporally smooth the physical fields after the EnKF update (Appendix A). Without this filtering, imbalances introduced by the EnKF update would have resulted in numerical noise during the forecast step, which in turn would have contaminated the forecasts and the covariance estimate during the next assimilation step, degrading the accuracy of the analysis. The digital filter was active for forecast hours 0–3, and was turned off for 3–6 hr. One effect of the digital filter switching on and off is an artificial positive trend in the tendency of precipitation rates from consecutive forecast windows (Figure 8a).

Instead of a digital filter, the version 3 system uses a 4D incremental analysis update (4DIAU) (Bloom *et al.*, 1996; Lei and Whitaker, 2016) to mitigate the imbalances introduced by the EnKF update. Essentially, the updates calculated by the EnKF analysis step are applied as a model forcing at every time step within the forecast, preventing gravity wave noise from contaminating its short-term evolution. Unlike the digital filter, the temporal smoothing in the 4DIAU is effectively only applied to analysis increments, not to the fields output from the model, thereby eliminating the spurious tendency trends seen during the forecasts of version 2c. Figure 8b

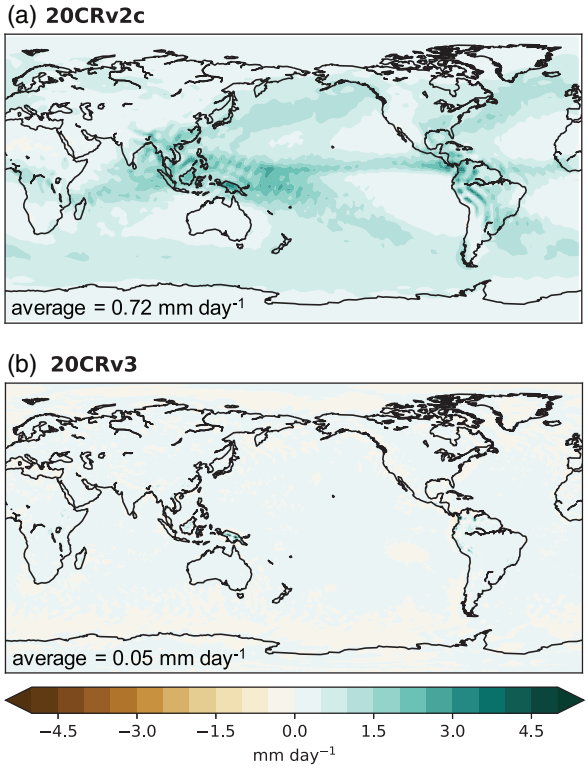

FIGURE 8 Maps of the year 2002 annually averaged differences (mm/day) between 3–6 hr forecasted precipitation rate and 0–3 hr forecasted precipitation rate in (a) 20CRv2c (which uses the digital filter) and (b) 20CRv3 (which uses the IAU instead of the digital filter). In a perfect system, these differences would be approximately zero

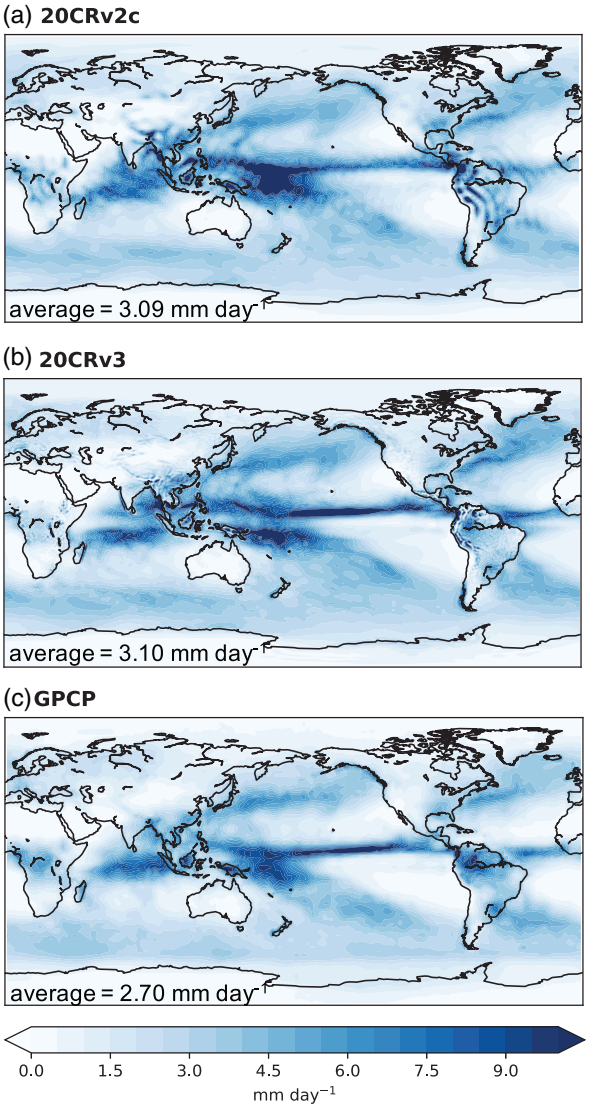

FIGURE 9 Maps of the year 2002 annually averaged forecasted precipitation rate (mm/day) in (a) 20CRv2c, (b) 20CRv3, and (c) the GPCP satellite/gauge blended fields

demonstrates the improvement over the digital filter: the precipitation rate biases have almost entirely disappeared. Note that the spatial average of the tendencies (difference between 3–6 and 0–3 hr forecast precipitation rates) from 20CRv2c fields (Figure 8a) is 0.72 mm/day, while the average from 20CRv3 (Figure 8b) is 0.05 mm/day. These figures show the annual average for 2002 but are representative of all available years.

Despite this change, the global annual average precipitation rate in tests with the 20CRv3 system is nearly the same as 20CRv2c. Figure 9 shows the 2002 annual average precipitation rates for (a) 20CRv2c, (b) 20CRv3, and (c) the gridded, blended satellite/gauge precipitation dataset from NASA's Global Precipitation Climatology Project (GPCP; Adler *et al.* 2003). The 20CRv3 field has a stronger separation in the western tropical rainband than 20CRv2c, leading to a double Inter-Tropical Convergence Zone (ITCZ) that is not as apparent in the GPCP dataset. Note that both versions of 20CR also overestimate global precipitation rates relative to GPCP; investigations into this issue are ongoing.

Figures 8a and 9a show another artifact of 20CRv2c, previously identified by Kent *et al.* (2013): namely, a spectral ringing characteristic in the precipitation mean and difference fields. This artifact is even more apparent in the 10 m wind fields (Figure 10a). This effect is the result of an error in the spectral transform of a higher-resolution orography to the lower-resolution field used as an input to 20CRv2c (Figure 10c). Since 20CRv3 uses a model at a higher resolution than 20CRv2c, it does not suffer from this issue as dramatically (Figure 10b, d), and the precipitation and wind fields lack any spectral ringing signals.

4 | ADDITIONAL IMPROVEMENTS IN 20CRV3

In addition to the specific developments in the version 3 system that were designed to ameliorate issues in 20CRv2c, early tests with the 20CRv3 system suggest that it will benefit from other updates to the system leading to overall improvements. In particular, the version 3 system uses adaptive

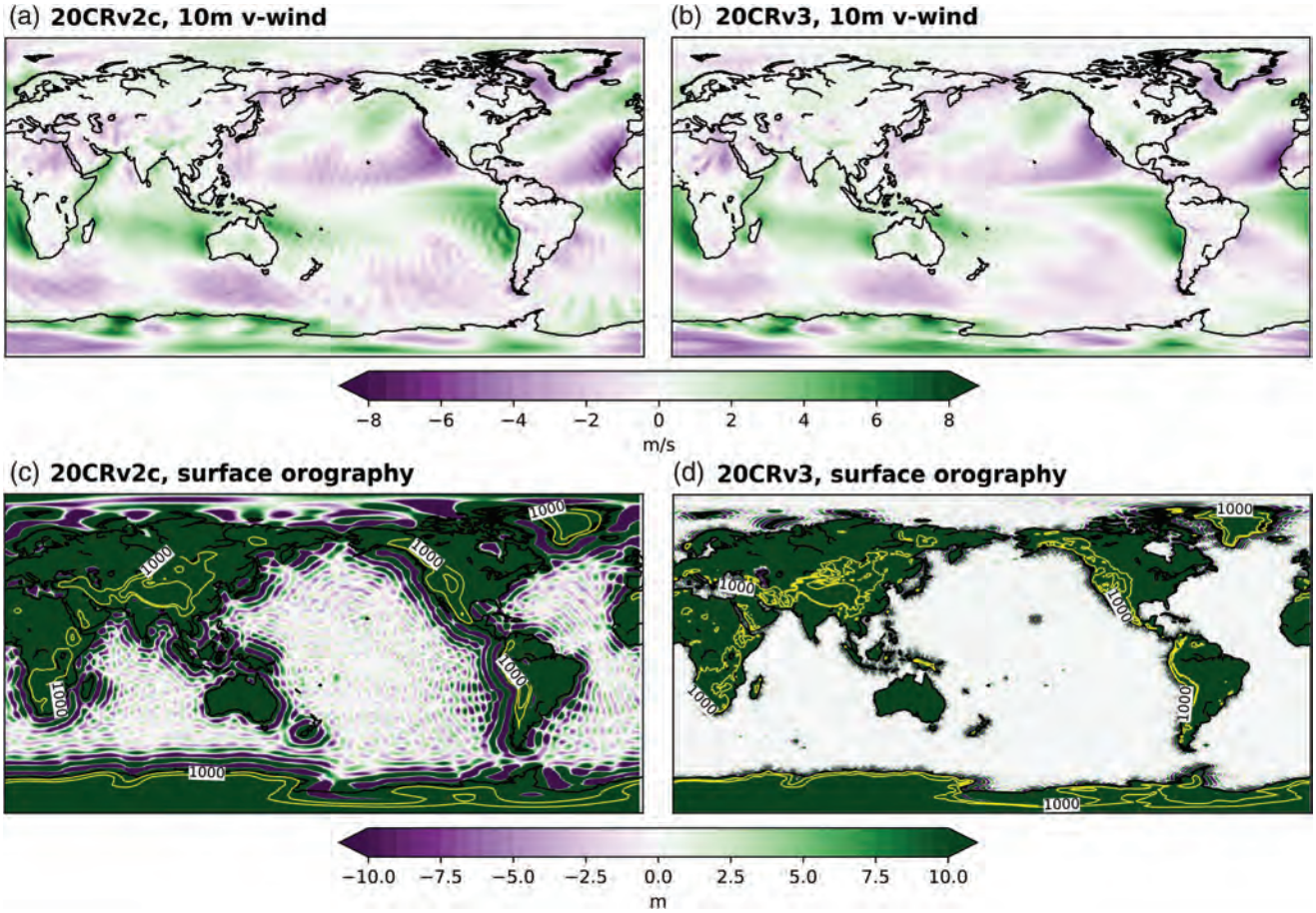

FIGURE 10 Maps of the year 2002 annual average of (a, b) 10 m meridional wind fields (m/s) and (c, d) surface orography fields (m) from (a, c) 20CRv2c and (b, d) 20CRv3. The 1,000 and 2,000 m contours have been plotted in yellow in (c) and (d), with the 1000 m contour labeled

QC and localization and specifies smaller tropical cyclone observation errors. These changes, in conjunction with a newer, higher-resolution forecast model, a larger observational database, and the improvements described in Section 3, yield results suggesting that 20CRv3 will have smaller forecast errors, large-scale reductions in model bias, and more accurate representations of hurricanes.

4.1 | Adaptive methods for assimilating observations

All versions of 20CR use an ensemble Kalman filter to assimilate observations. It is well-known that localization is required to prevent spurious ensemble cross-correlations from developing far away from the assimilated observations (Houtekamer and Mitchell, 1998; 2001; Hamill *et al.*, 2001; Anderson, 2007). The use of localization in ensemble DA systems for weather prediction is crucial, since current-generation systems are being run with ensemble sizes many orders of magnitude smaller than the size of the forecast model state vector. Traditional localization schemes use a smooth cut-off function, such as the piecewise continuous fifth-order polynomial function described by Gaspari and Cohn (1999), to taper the covariances to zero at a given distance away from an observation. Typically, this Gaspari–Cohn localization is a function of only the horizontal distance between an observation and a state variable, and is described by a single parameter that is related to the distance at which the Gaspari–Cohn function goes to zero. The optimal value for the localization length-scale may be a function of many aspects of the DA system, such as the density of the observing system and the scale of the phenomena being observed. This makes tuning the localization length-scale difficult, especially when the observing system is very inhomogeneous, and many different scales are being observed simultaneously. The 20CRv2c system used a localization radius of 4,000 km for all times and locations based on early tests (Whitaker *et al.*, 2004); a relatively large value was chosen to maximize the use of observations in data-sparse regions and to minimize the generation of small-scale noise by the EnKF update. In addition, a five-step QC process was employed in 20CRv2c; this is the same process used in 20CRv2 and described in Appendix B of Compo *et al.* (2011).

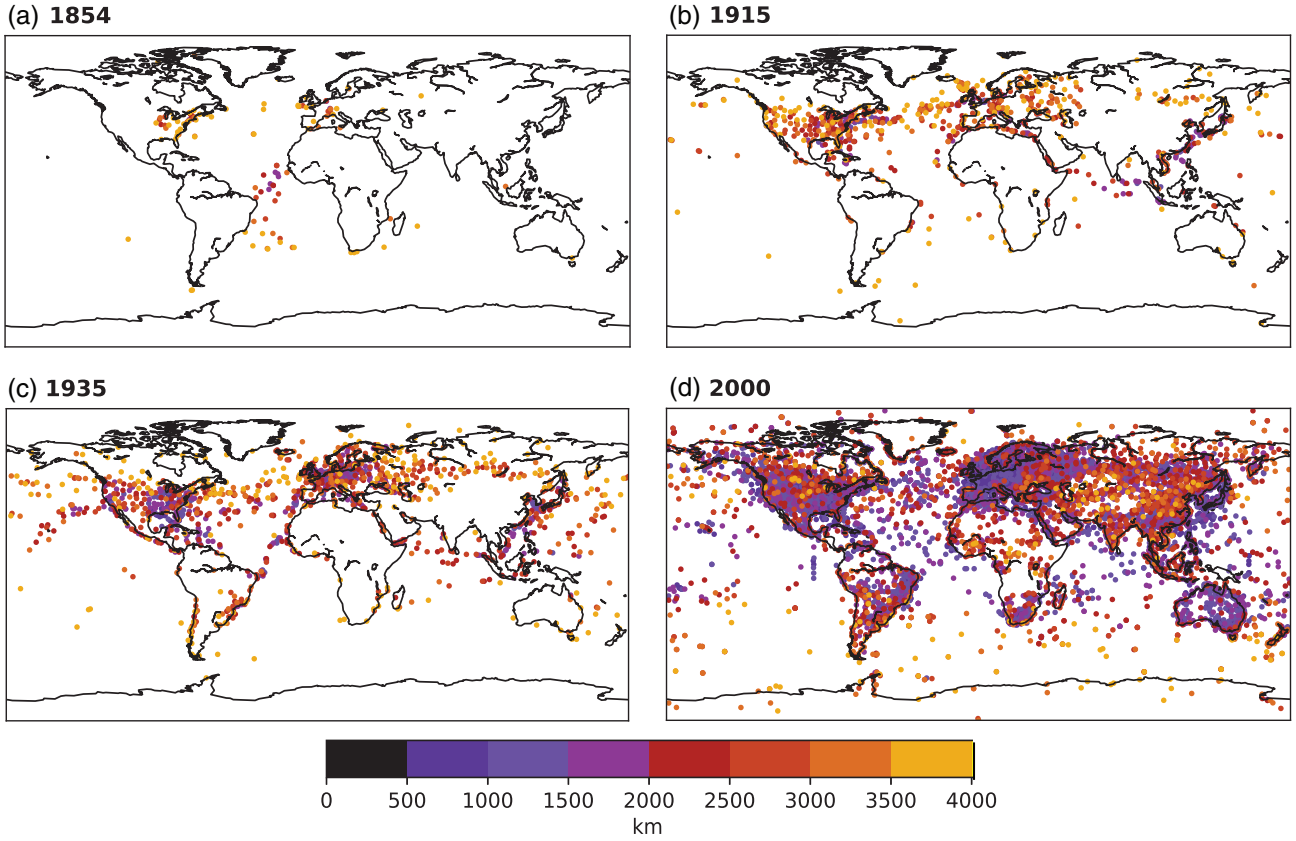

FIGURE 11 Localization values (km) for all observations assimilated in 20CRv3 at 1200 UTC on 1 June for the years (a) 1854, (b) 1915, (c) 1935, and (d) 2000. Note that comparable plots for 20CRv2c would consist entirely of light orange circles (localization value of 4,000 km)

Since the 20CR system only assimilates surface pressure observations and the network can become quite sparse in the 19th century, it is important to extract the most information from each observation. In order to make better use of the observations, 20CRv3 uses an adaptive QC procedure jointly with an adaptive localization algorithm. Observations must pass two initial gross QC steps: if the observation is outside a plausible range or if the observation is too far from the first guess, the observation is rejected. The first step will reject an observation that is outside the range of 850 to 1090 hPa. The second step will reject an observation x_{ob} if it is too far from the first guess x_b:

$$|x_{ob} - x_b| > 3.2\sqrt{\sigma_b^2 + \sigma_{ob}^2}, \tag{4}$$

where σ_b^2 is the variance of the first-guess ensemble interpolated to the observation time and location and σ_{ob}^2 is the observation error variance. Unlike 20CRv2c, version 3 does not utilize a "buddy check" or a thinning algorithm to reject observations which degrade the fit of the analysis to nearby observations or which do not decrease the analysis ensemble spread. Instead, the adaptive QC assigns these observations larger errors and smaller localization radii, so that their region of influence is essentially zero. Details of the adaptive QC and localization procedures used in 20CRv3 can be found in Appendices C and D.

Figure 11 shows maps of observations within a single assimilation window for four test years: 1854, 1915, 1935, and 2000. Note that, as the observation network becomes denser, the localization lengths generally decrease. In 1854 and 1915, the observation network is relatively sparse, and most observations have localization length-scales near the maximum allowed of 4,000 km. In the year 2000, however, most observations located within densely observed areas have localization length-scales closer to 1000 km, though there are a few observations within these areas that the algorithm selects for greater localization length-scales. Observations located within areas that are sparsely observed (such as the Southern Ocean and Antarctica) still have fairly large localization length-scales in the year 2000. This new procedure allows many more observations to be assimilated within 20CRv3 while adaptively allowing observations with significant beneficial effects to have larger ranges of impact, and observations with less beneficial effects to have smaller ranges.

4.2 | Observation statistics

Observations have a large impact on overall performance of reanalyses: inaccurate observations as well as the incorrect

prescription of their errors can impact global fields and their trends (recall the global SLP bias in Figure 5). It is crucial, then, to investigate the behaviour of their statistics in the context of the full system. Here we show that, while 20CRv2c performs fairly well under many measures, the updated algorithms used in 20CRv3 produce clear improvements in several test periods.

Statistics of the departures of observations from the first-guess field can provide one measure of how well the entire system is performing, particularly when compared with statistics of the expected errors. The "actual" error is defined as:

$$\mathrm{RMSE}_{\mathrm{actual}} = \left[\frac{1}{N_{\mathrm{obs}}} \sum_{i=1}^{N_{\mathrm{obs}}} \left(x_{\mathrm{ob},i} - x_{\mathrm{b},i} \right)^2 \right]^{1/2}, \qquad (5)$$

where i indexes all N_{obs} observations that are contributing to the statistic (in space or time), $x_{\mathrm{ob},i}$ is the ith observation, and $x_{\mathrm{b},i}$ is the first-guess field interpolated to the ith observation time and location. The expected error is then defined as:

$$\mathrm{RMSE}_{\mathrm{exp}} = \left[\frac{1}{N_{\mathrm{obs}}} \sum_{i=1}^{N_{\mathrm{obs}}} \left(\sigma^2_{\mathrm{ob},i} + \sigma^2_{\mathrm{b},i} \right) \right]^{1/2}, \qquad (6)$$

where i indexes the same N_{obs} observations as above, $\sigma^2_{\mathrm{ob},i}$ is the ith observation error variance, and $\sigma^2_{\mathrm{b},i}$ is the variance of the first-guess ensemble, interpolated to the ith observation time and location. As shown by Desroziers *et al.* (2005), under the assumptions that the observation and background errors are uncorrelated and unbiased, these errors should be equivalent. In a DA system that is performing well, the actual errors should not be larger than the expected errors. We consider time series of regionally and annually averaged surface pressure forecast errors for the Northern Hemisphere (20–90°N), Tropics (20°S to 20°N), and Southern Hemisphere (90–20°S). That is, the statistics in Equations 5–6 are calculated using all observations in the given region for each year. These are plotted in Figure 12 for 20CRv2c (blue) for all years and 20CRv3 (black) for three test periods.

As expected from EnKF theory and seen with 20CRv2 (Compo *et al.*, 2011), errors decrease in time as observations are denser and more accurate. Note that this is not a result of any "tuning". The Southern Hemisphere errors from 20CRv2c match the expected errors particularly well in all decades after 1860. The 20CRv2c errors in the Tropics are less consistent, and the actual errors in the Northern Hemisphere are almost always larger than the expected errors, by over 1 hPa in earlier decades and by 0.3–0.5 hPa in recent decades. This suggests that, in the 20CRv2c system, the errors assigned to observations might be too low, the first-guess ensemble spread is too small, the observations or first-guess fields are biased, or a combination of these.

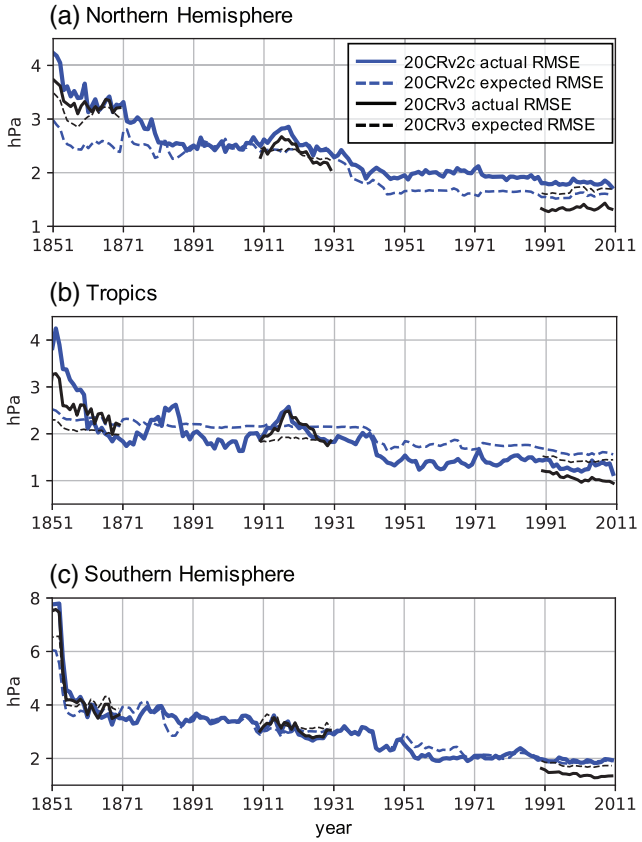

FIGURE 12 Time series of actual (solid) and expected (dashed) annual first-guess RMS errors for observations assimilated in 20CRv2c (blue) and 20CRv3 (black) averaged over (a) the Northern Hemisphere (20 to 90°N), (b) the Tropics (20°S to 20°N), and (c) the Southern Hemisphere (90 to 20°S)

The preliminary 20CRv3 errors (black curves in Figure 12) show improvements in nearly every test period. While the actual errors in the 1851–1870 test period are still larger than the expected errors, this discrepancy is smaller than in 20CRv2c. Recall from Section 3 that, prior to 1871, a bias correction was made to marine observations in 20CRv3, in addition to the station observation bias correction. Although dry air mass conservation is mainly responsible for removing the low-pressure bias (Figure 7), the ship bias correction provides further improvement. For the modern period in all regions, the actual error is always less than expected. This suggests that the assigned observation errors may be too large for modern years, and future work may investigate the effects of time-varying observation errors as in Laloyaux *et al.* (2018). Nevertheless, the overall effect of the new algorithms in the version 3 system, including ship bias correction, is to decrease the RMS errors in several different test periods.

These improvements in RMS errors may be due to the updated model, as well as to the new algorithms implemented in the 20CRv3 system. This is supported by investigations into the station bias corrections. Recall that these corrections are based on 60-day average differences between observations

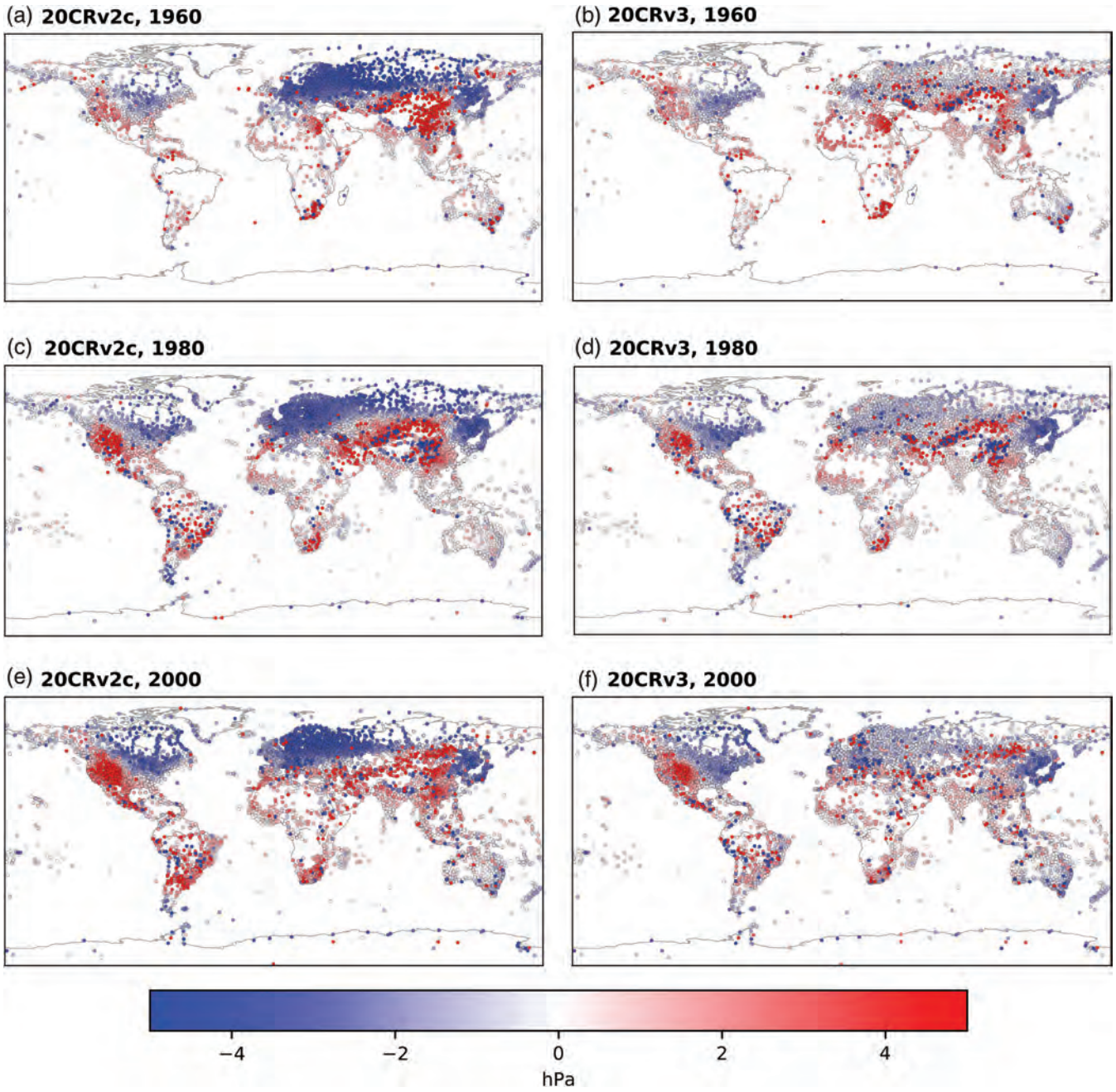

FIGURE 13 Annual average station pressure biases (hPa) for (a, b) 1960, (c, d) 1980, and (e, f) 2000, calculated from (a, c, e) 20CRv2c and (b, d, f) 20CRv3

from each station and the first-guess pressure at that location. Ideally, the consistent mismatches between observations and the model first guess are biases in the observations that are removed; however, a mismatch could be due to model errors, and the algorithm would actually be adjusting the observations away from reality and towards the biased model state. In particular, it is likely that biases with large-scale spatial patterns are model errors, though they could result from national issues producing similar biases (Slonosky and Graham, 2005) or international changes in observing practices that are adopted by nations at different rates (discussion in Trenberth and Paolino 1980). Conversely, small-scale biases may point to a misassignment of station elevation or position: these are observational biases that should be corrected.

Figure 13a,c,e shows the annual average station pressure biases from 1960, 1980, and 2000 that were removed from observations in 20CRv2c. Note the annual values in the region over eastern Europe and much of northern Asia are consistently negative, suggesting a model bias (also van den Besselaar *et al.* 2011). Conversely, the version 3 data does not show the same spatial or temporal consistency of negative values in this region (Figure 13b,d,f). This suggests that the model used in version 3 may be less biased than in 20CRv2c.

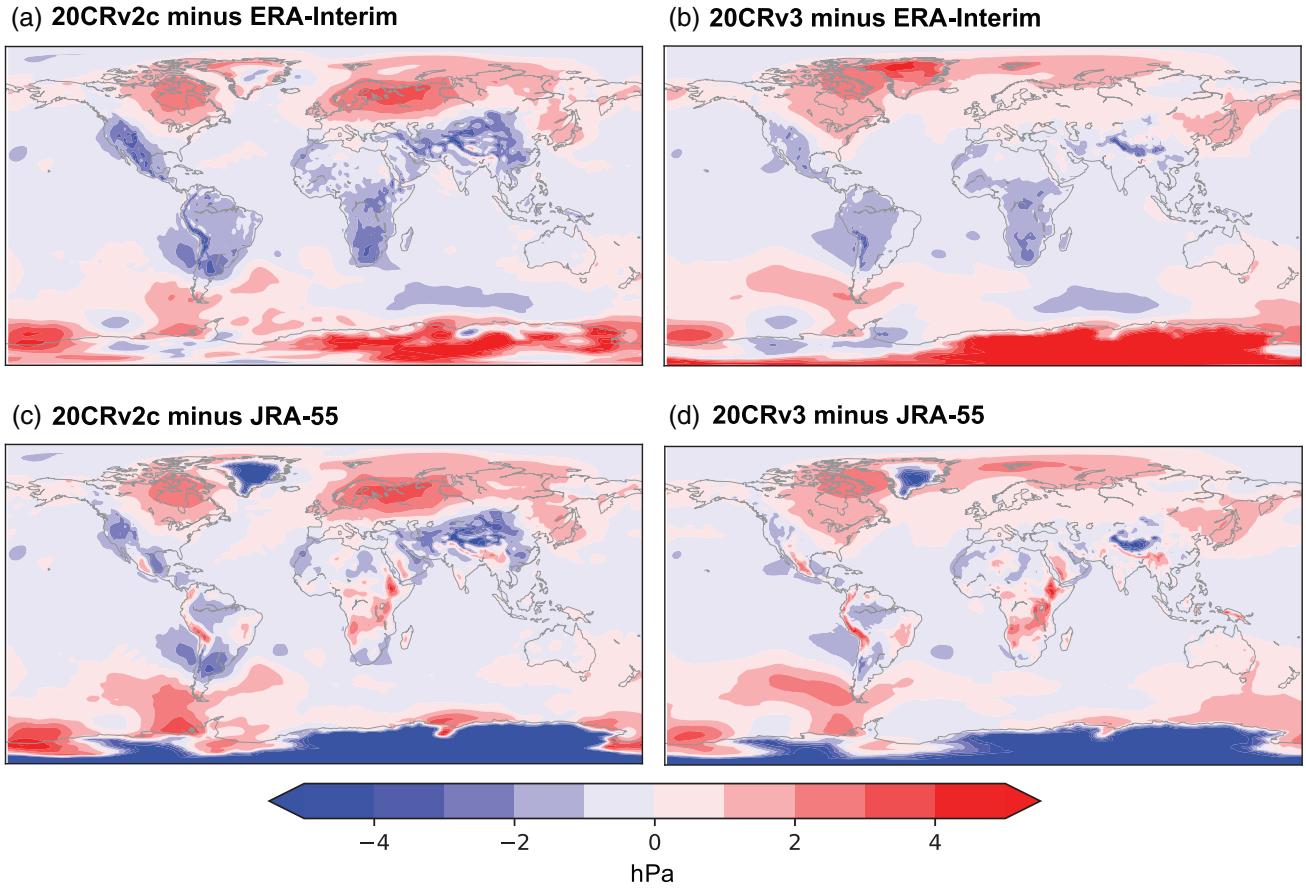

FIGURE 14 Maps of the year 2000 annually averaged sea-level pressure differences (hPa) between (a) 20CRv2c and ERA-Interim, (b) 20CRv3 and ERA-Interim, (c) 20CRv2c and JRA-55, and (d) 20CRv3 and JRA-55, all plotted on approximately a 1.5° by 1.25° grid

Another cause could be due to the process of the station bias correction: 20CRv2c calculated biases from the observations interpolated to the model surface, but inadvertently applied the correction at the level of the observation; in 20CRv3, the bias correction is calculated and applied at the level of the observation.

Unlike the Eurasian biases, consistent regional biases over the US have hardly changed in 20CRv3; it is unclear whether this is a model bias effect or not. Other possible causes of large-scale biases include orography (biases over mountain ranges tend to be consistent; e.g. the Himalaya) and nationality (due to country-specific calibration and correction methods; cf. the Canadian "50-foot rule" (Slonosky and Graham, 2005). Finally, version 3 includes many other changes to the assimilation method and the observation handling, as well as changes to the forecast model, so it is difficult to confidently conclude that the negative bias over northern Asia was a model error that has been fixed. Regardless, the overall mismatch between station observations and first-guess fields has been notably diminished in 20CRv3. This leads to analyzed SLP fields (Figure 14) and 500 hPa geopotential height fields (Figure 15) that are closer to those of ERA-Interim and JRA-55, particularly in northern Eurasia.

4.3 | Representation of hurricanes

Historical reanalyses are especially useful for studying extreme weather events, since these are by definition rare but high-impact events. Tropical and extratropical cyclones are of particular interest as they can result in loss of life and enormous financial costs. To improve understanding and predictions of these storms, it is necessary to improve our understanding of the large-scale drivers of them as well as how storm characteristics are changing as the climate changes. If historical reanalyses can accurately capture these storms, they provide a long, consistent sample of such extreme events and their associated large-scale environment.

In order to represent tropical cyclones (TCs), 20CRv2c assimilated TC data from IBTrACS (Knapp *et al.*, 2010) in addition to land station and marine pressure observations. IBTrACS consists of actual pressure measurements, pressure reports calculated as time-interpolated values from tropical cyclone warning centers, and wind-derived central pressure reports (Knapp *et al.* 2010 and Compo *et al.* 2011 give more details). Since these data are often significantly lower pressure values than the nearby station observations, they would generally fail the QC step that compares observations to their nearest neighbours (Appendix B of Compo *et al.*, 2011);

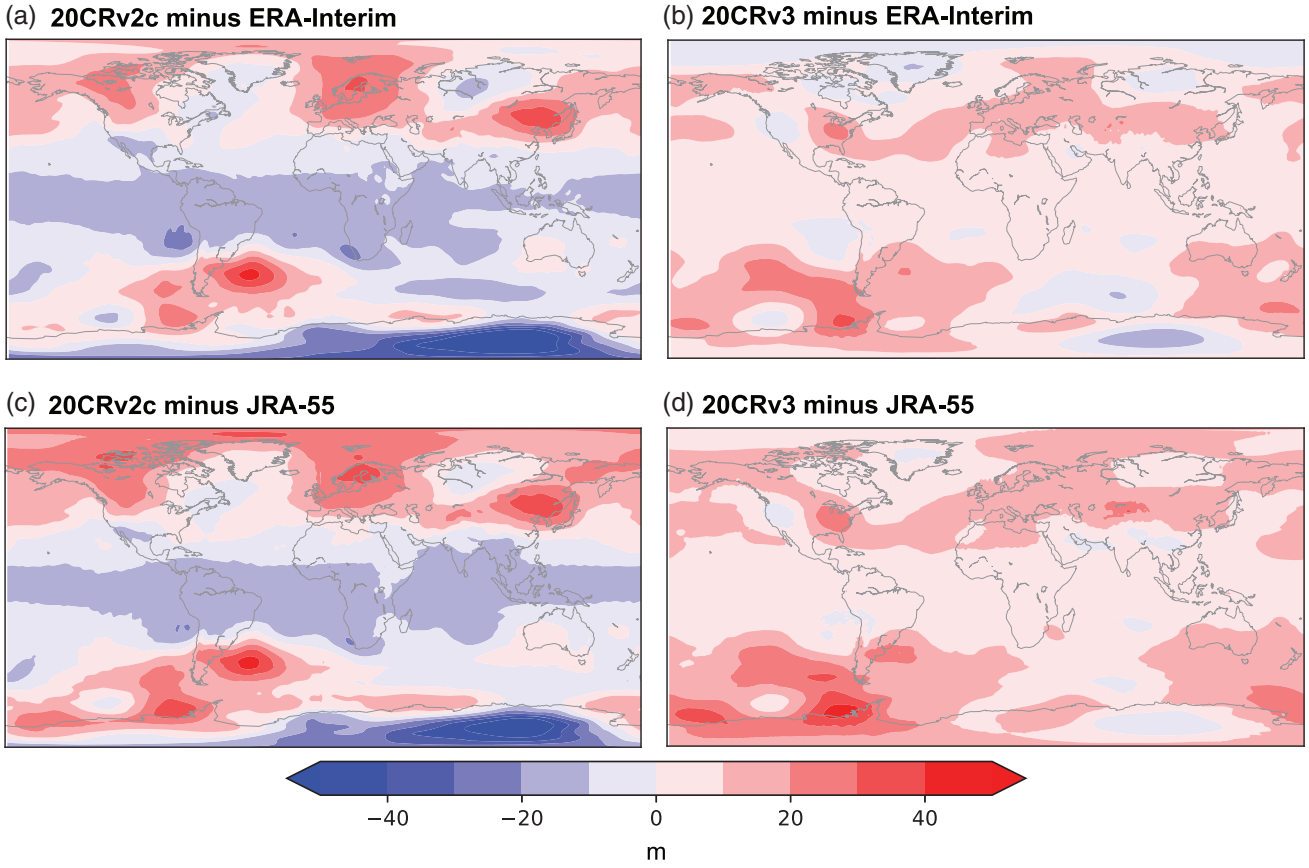

FIGURE 15 Maps of the year 2000 annually averaged 500 hPa geopotential height differences (m) between (a) 20CRv2c and ERA-Interim, (b) 20CRv3 and ERA-Interim, (c) 20CRv2c and JRA-55, and (d) 20CRv3 and JRA-55, all plotted on approximately a 1.5° by 1.25° grid

therefore, the 20CR system has IBTrACS data bypass these checks and assimilates these deep-low data.

In version 2c, these data were assigned observation errors that were much higher than for any other type of observation to prevent numerical instabilities from arising immediately after assimilation (Table A1 in Appendix). Despite digital filtering to smooth the evolution of the post-update fields, tests using smaller errors would occasionally still generate amplifying gravity waves and numerical instability. While the large error assignment eliminated this problem, the resulting 20CRv2c analyses can have central pressure values that are much higher than the IBTrACS data, sometimes by 40 hPa or more. The version 3 system, with an updated, higher-resolution forecast model and 4DIAU, can use these observations more effectively. Assimilating TC low-pressure values into the version 3 system does not generate instabilities, and so the IBTrACS data can be assigned smaller observation errors (Table 1). This often yields stronger cyclones with central pressure analyses that are closer to the original IBTrACS value.

As an example, we investigate a strong hurricane that hit Galveston, Texas in August 1915. Figure 16 illustrates the analyzed SLP fields (contours) from four reanalyses, as well as locations of observations available to each system between 16 August 1915, 2100 UTC and 17 August 1915, 0900 UTC (circles); this window was chosen so that observations assimilated at 0000 UTC would be shown as well as those assimilated at 0600 UTC. For each system, observations that were assimilated are shown as solid circles, while observations that were rejected by the system's QC step are open circles. Blue circles represent station and marine observations, and red circles represent IBTrACS data. 20CRv2c (Figure 16a) assimilated the IBTrACS report of 940 hPa on 17 August 1915 at 0600 UTC, producing an analyzed value of 986 hPa at the centre of the storm. In version 3 (Figure 16b), the storm is even stronger, and the analyzed value at the centre of the storm has decreased to 962 hPa, reducing the observation–analysis departure calculated from 20CRv2c by half. Since the IBTrACS reports were available and assimilated into 20CRv2c and 20CRv3 beginning on 1 August 1915, this hurricane evolved and strengthened continuously in time in both versions of 20CR (not shown.) For comparison, the ERA-20C QC rejected the IBTrACS observations (Figure 16c) and analyzed a low pressure system that is weaker than that of 20CRv2c and has a misaligned centre. The CERA-20C system also rejected the IBTrACS observations, but additionally assigned larger errors to the nearby station data than ERA-20C (Laloyaux *et al.*, 2018), thereby showing

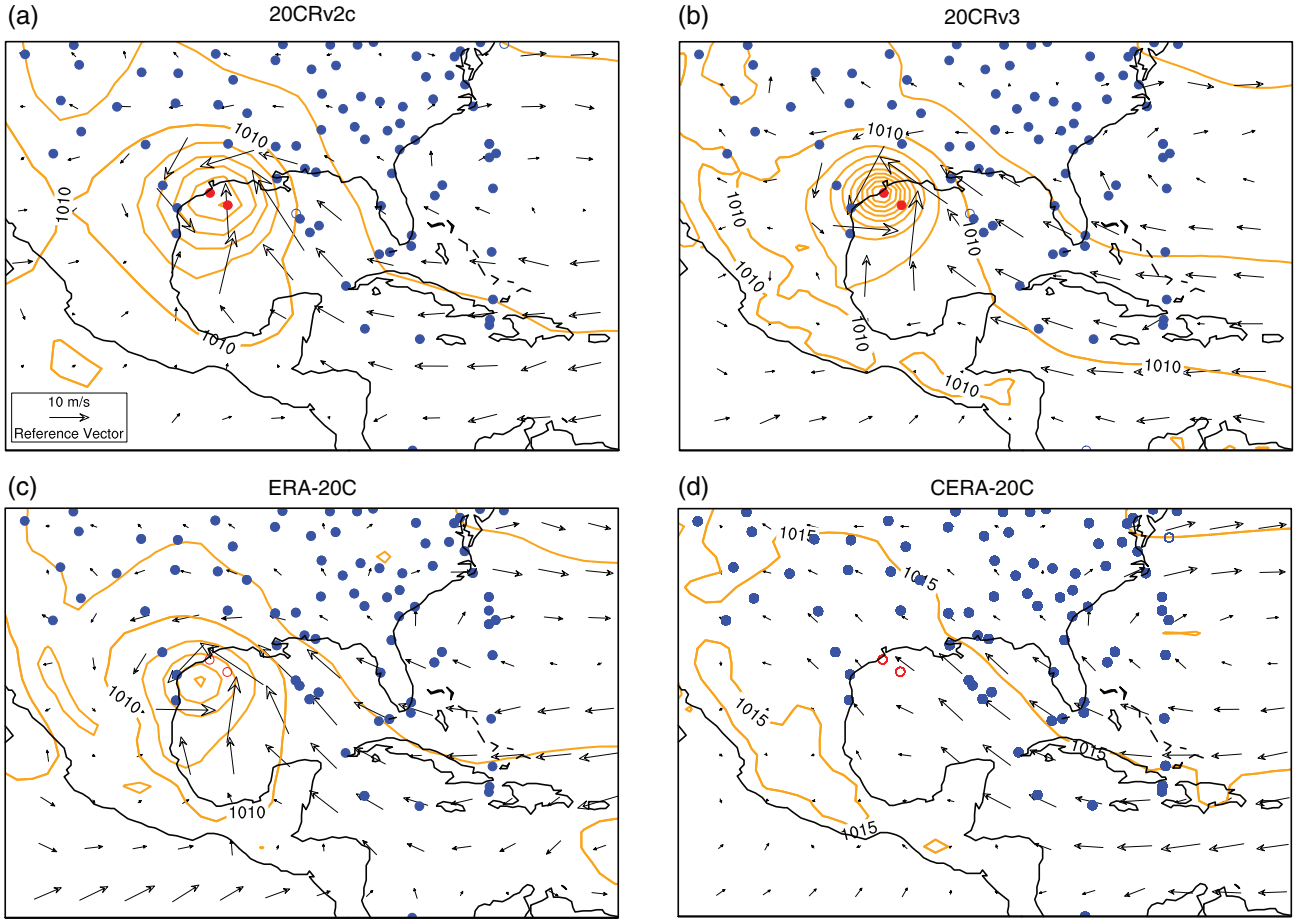

FIGURE 16 Sea-level pressure (orange contours with interval 5 hPa) and wind fields (vectors; m/s) for the 1915 Galveston hurricane, 17 Aug 1915 0600 UTC, from (a) 20CRv2c, (b) 20CRv3, (c) ERA-20C, and (d) CERA-20C. Locations of available observations taken between 16 August 1915 2100 UTC and 17 August 1915 0900 UTC are shown by circles: station and marine observations are blue, IBTrACS data are red, solid circles denote observations that were assimilated, and open circles denote observations that were rejected

no trace of the storm (Figure 16d). Laloyaux *et al.* (2018) conducted experiments with the CERA-20C system in which this type of observation was white-listed, and found this yielded better performance than CERA-20C for two hurricanes in 1900 and 2005 (their Figure 5).

5 | REMAINING ISSUES

While the changes from the version 2c system to the version 3 system have resulted in many improvements across the board, there are several remaining issues in 20CRv3 as well as new questions that have arisen. For instance, recall (Figure 9) the precipitation biases that have strengthened in 20CRv3, particularly the appearance of a double ITCZ and the overestimation of global precipitation rates relative to GPCP. The Southern Hemisphere confidence fields (Figure 4) also demonstrate that there is some remaining large uncertainty over Antarctica (though these areas are relatively small), despite tests that led to adjusting the inflation parameter in the Southern Hemisphere.

Figures 7 and 14a,b demonstrate another potential issue with the Southern Hemisphere: a high pressure bias over Antarctica. Figure 7 shows that the 1854 annual average SLP over Antarctica in 20CRv3 is several hPa higher than the 20CRv2c modern climatology, and this anomaly is larger in 20CRv3 than it was in 20CRv2c. Figure 14 demonstrates that 20CRv3 also displays this high pressure bias in a modern difference calculated with respect to ERA-Interim. However, the strong difference relative to ERA-Interim is mainly over the Antarctic landmass, which has a fairly high topography, so the SLP field is likely not an appropriate variable to consider when diagnosing the mass or circulation field of this region. Indeed, the SLP difference with JRA-55 (Figure 14c,d) has the opposite sign in this region.

A third Southern Hemisphere issue, regarding a trend in sub-Antarctic SLP, was first brought to light during an investigation of ERA-20C (Poli *et al.* 2015,2016) and is demonstrated in Figure 17. This figure shows the seasonal time series of SLP area-averaged poleward of 60°S for 20CRv2c (blue), ERA-20C (green), CERA-20C (gold), and ERA-Interim

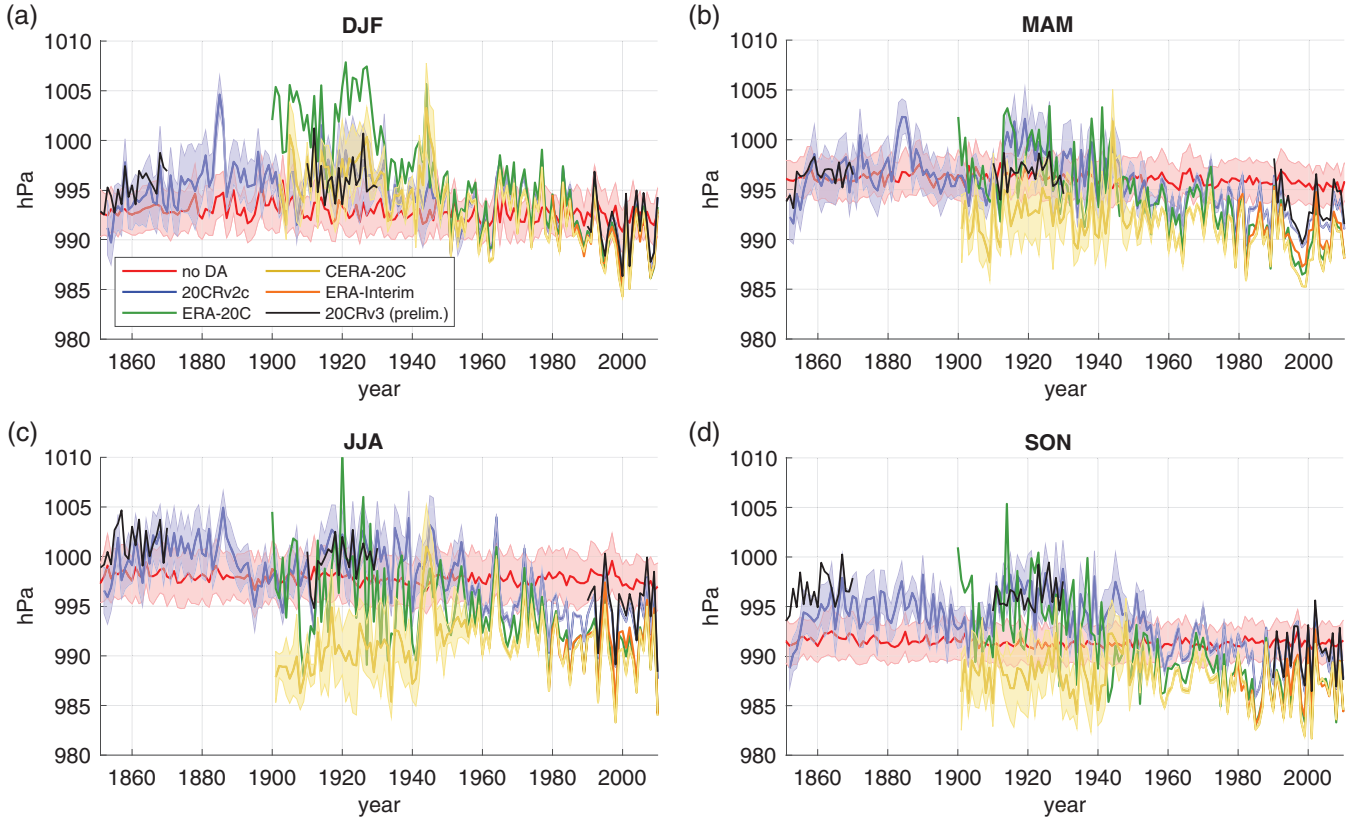

FIGURE 17 Time series of seasonal sea-level pressure poleward of 60°S from 20CRv2c (blue), ERA-20C (green), CERA-20C (gold), ERA-Interim (orange), a non-assimilating model simulation using the 20CRv2c system (red), and preliminary 20CRv3 data without confidence intervals (black). Shading represents one ensemble standard deviation when available

(orange), as well as data from an ensemble of model simulations using the 20CRv2c system but which did not assimilate any observations ('no DA', red) and preliminary 20CRv3 data for the test periods 1851–1870, 1910–1930, and 1990–2010 (black). ERA-20C has a high pressure bias south of 60°S in the early 20th century that is particularly strong in austral summer (green curve in Figure 17a between 1900 and 1930). Comparisons with 20CRv2c show that it has a similar bias as ERA-20C in other seasons, but a weaker bias than ERA-20C in Dec-Jan-Feb. There is a drop and subsequent increase in SLP from 20CRv2c in all seasons (most notably in Sept-Oct-Nov) from 1890 to 1910, with another significant drop-off between 1940 and 1960. The preliminary 20CRv3 data agree with the 20CRv2c data for the most part, though the early 20th century March-April-May SLP has been diminished. In all seasons though, the modern 20CRv3 SLP is still about 5 hPa lower than the early 20th century SLP.

ERA-20C and the 20CR datasets used entirely different models and assimilation methods but show similar trends, which suggests the culprit is in the observations. Poli *et al.* (2015) and Laloyaux *et al.* (2018) assert that the problem was caused not by a bias in the observations, but by the spatial pattern of observations at this time in the Southern Hemisphere. In particular, most of the observations are located in the subtropical high-pressure belt; the positive increments from assimilating these observations were communicated to the unobserved, and thus unconstrained, region farther south. They argue that this is caused by observation errors that are too small in ERA-20C; thus in CERA-20C, larger observation errors were assigned in this time period. This has significantly decreased the trend in SLP from CERA-20C in all seasons (gold curves in Figure 17), though it remains somewhat in Dec-Jan-Feb, when the bias in ERA-20C was most obvious. Hegerl *et al.* (2018) point out the HadCRUT4 temperature dataset (Morice *et al.*, 2012) also exhibits anomalously low SSTs during the years 1906–1915 in the Southern Ocean (their Figure 2c), possibly due to instrumental biases or uncertainties in the sea ice fields; investigations into this issue and possible connections with the SLP signals in Figure 17 are ongoing.

6 | CONCLUSIONS

With the growing need to understand and predict climate and extreme weather variations on decadal to centennial time-scales, the use of historical reanalyses continues to expand in areas such as assessments of long-term climate change, investigations of extreme events, and detailed

histories of weather. Therefore it is becoming more important that these reanalyses be reliable, both in their state estimates and their quantification of uncertainty. Users must recognize when and where historical reanalyses can be confidently utilized, and when caution should be taken (or a different dataset chosen). This work seeks to illuminate particular aspects of 20CRv2c that require careful consideration, the ways in which these issues informed the development of the 20CRv3 system, and particular aspects of 20CRv3 that show preliminary improvements over 20CRv2c.

The Twentieth Century Reanalysis version 2c improved upon several issues discovered in the previous NOAA-CIRES historical reanalysis, 20CRv2, but other problems remained. They provided specific focus areas when developing the NOAA-CIRES-DOE 20CRv3 system. Indeed, many of the issues in 20CRv2c discussed here have been ameliorated in 20CRv3 due to a combination of factors: a newer NCEP GFS forecast model with higher resolution; improved DA algorithms, observation processing, and QC; and an updated ISPD observation database. Several other issues with 20CRv2c exist that have not been discussed here, including spin-up effects in sea ice thickness, snow depth, and soil moisture, and biases in the upper-stratospheric temperatures; some of these issues are reduced in 20CRv3 and will be discussed in future work.

Preliminary results with the 20CRv3 dataset are quite promising, though they are already highlighting areas for future research, particularly in the "Deep South" of the Southern Hemisphere. The confidence in that region remains too low; further work regarding the relaxation-to-prior-spread inflation algorithm in this region may be necessary to increase the confidence to more realistic values. A larger set of available observations in this region would also increase the confidence (recall Figure 2), motivating greater data rescue efforts here. Gathering high-quality observations in these sparsely observed regions remains a challenge within the data rescue community (Allan *et al.*, 2011; Brönnimann *et al.*, 2018), but new data rescue efforts (SouthernWeatherDiscovery.org) are beginning to address this. More data are also needed in other sparsely observed regions, as well as globally in the early 19th century. Other data rescue efforts (including ACRE activities, WeatherRescue.org, the Copernicus Climate Change Service South America data rescue project, and the UK/China Climate Science for Service Partnership) have the potential to significantly add to the observational database in these regions.

Despite some remaining challenges with 20CRv3, there are early suggestions that this dataset will be useful for studies in which 20CRv2c required more cautious analysis: for example, tropical cyclones seem to show much stronger signals in 20CRv3 than in 20CRv2c. This suggests that 20CRv3 may be used for validating ongoing historical tropical cyclone research that extends IBTrACS back in time (Diamond *et al.*, 2012), and for corroborating partial or discontinuous storm track information (e.g. when storm systems passed close to islands or ships). Utilizing an updated inflation algorithm also allows for more consistent studies of long-term trends and uncertainty, where 20CRv2c exhibited artificial signals due to abrupt parameter changes.

Since the process of creating historical reanalyses is a continuous cycle of improvement, we are already looking ahead to further upgrades to the 20CR system. In particular, NCEP has recently significantly updated their global forecast system with a finite volume, cubed sphere model (Harris and Lin, 2013) (preliminary documentation available at https://vlab.ncep.noaa.gov/web/fv3gfs/; accessed 7 July 2019); the changes resulting from this model versus the previous spectral model need to be investigated. Recent investigations into coupled DA algorithms, and the first implementation of a quasi-strongly coupled DA algorithm in CERA-20C (Laloyaux *et al.*, 2018), suggest that future versions of 20CR could benefit from coupled systems. Finally, while all versions of 20CR so far have only assimilated surface pressure, the possibility of assimilating other types of data (such as marine winds, sea ice observations, or precipitation) seems to be more feasible as DA algorithms continue to mature.

ACKNOWLEDGEMENTS

Support for the Twentieth Century Reanalysis Project is provided by the Physical Sciences Division of the NOAA Earth System Research Laboratory, the US Department of Energy, Office of Science, Office of Biological and Environmental Research (BER), and by the National Oceanic and Atmospheric Administration (NOAA) Climate Program Office. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under contract no. DE-AC02-05CH11231. Some computing was also performed on NOAA's Remotely Deployed High Performance Computing Systems.

P. Brohan's (UK Met Office) extensive work on ICOADS and on the marine bias correction algorithm used in 20CRv3 is gratefully acknowledged. M. Benoy for the Citizen Science Unit of the Australian Meteorological Association, working with the Australian Bureau of Meteorology, is gratefully acknowledged for ongoing support. Collaborations with N. Rayner and the UK Met Office in the development, production, and use of HadISST boundary conditions are gratefully acknowledged. H. Maechel's (DWD) contribution of German climate observations is gratefully acknowledged. The authors would like to acknowledge the consultants of NERSC for their help with implementing the 20CRv3 system. The authors would also like to thank the following individuals for their invaluable contributions of observations to the ISPD: L. Alexander (University of New

South Wales), M. Barriendos (University of Barcelona), T. Brandsma (KNMI), Y. Brugnara (University of Bern), O. Bulygina (All-Russia Research Institute of Hydrometeorological Information), A. Dawson (University of Aberdeen), J. Filipiak (University of Gdansk), P. Groisman (NC State University Research Scholar at NOAA NCEI), J. Holopainen (University of Helsinki), D. Jones (Australian Bureau of Meteorology), T. Jónsson (Icelandic Met Office), J. A. López (Instituto Nacional de Meteorología, Madrid), O. Mestre (Météo-France), A. Moberg (Stockholm University), Ø. Nordli (Norwegian Meteorological Institute), M. Rodwell (ECMWF), T. Schmith (Danish Meteorological Institute), L. Srnec (Croatian Meteorological and Hydrological Service), M. Tolstykh (Hydrometcentre of Russia), N. Westcott (Midwestern Regional Climate Center), and P. Woodworth (National Oceanography Centre, UK). IBTrACS data are courtesy of K. Knapp (NOAA/NCEI). P. Laloyaux (ECMWF) is thanked for discussions regarding tropical cyclone representation. The efforts of the NCAR Data Engineering and Curation section, especially D. Schuster, R. Conroy, and C.-F. Shih are acknowledged. The technical support of the IT group of the NOAA Earth System Research Laboratory Physical Sciences Division is acknowledged. Partial support from SwissRe and collaboration with P. Zimmerli is acknowledged. Comments from A. Shlyaeva (CIRES/NOAA) and two anonymous reviewers improved this manuscript.

J. Kennedy was supported by the Met Office Hadley Centre Climate Programme funded by BEIS and DEFRA. M. Brunet acknowledges Copernicus Climate Change Service (C3S)/Data Rescue Service (DRS) (contract code: ATT 3670; FENIX: T17088S). J. Gergis was funded by Australian Research Council Project DE130100668. E. Hawkins was supported by the UK National Centre for Atmosphere Science. A. Kaplan acknowledges the 2002 grant from the LDEO Climate Center and NOAA awards NA03OAR4320179 and NA17OAR4310156. The research work of R. Przybylak and P. Wyszyński was supported by the National Science Centre, Poland (grant nos. DEC-2012/07/B/ST10/04002 and 2015/19/B/ST10/02933). M. A. Valente acknowledges Instituto Dom Luiz – Faculty of Science of the University of Lisbon, Instituto Português do Mar e da Atmosfera (IPMA), Instituto Geofísico do Porto, Instituto Geofísico da Universidade de Coimbra, projects SIGN (FCT-POCTI), ERA-CLIM (FP7) and ERA-CLIM2 (FP7).

The Justus Liebig University of Giessen, Germany is thanked for financial support to digitise, QC and analyse early instrumental meteorological data across the world. The following people are financially supported by the University of Giessen and digitised subdaily pressure data: L. Dergianli, G. Kelly, D. Xoplaki, V. Iakovoglou, E. Kaimasidou, E. Tsalkitzidou, M. Athanasiou, L. Ferger, A. Megalou, C. Chandolia, E. Fleitmann, P. Zafeiropoulou, N. Kong, M. Ostheimer, S. Dafka, K. Pometti, N. Gouta, P. Katsaouni, Z. Theorchari, A. Theocharis, F. Williams, J. Schermuly, C. Mett, L. Theile, J. Zuckermann, P. Strehlau, C. Samaras, A. Tsikerdekis, C. Athanasiou, J. Braun, T. Sperzel, J. Damster, N. Luther, M. Miltscheff-Petroff, M. Kelbling, D. Griechbaum, M. Hänsgen and J. Viezens. L. Dergianli is acknowledged for QC, coordination, management and preparation of all University of Giessen data. S. Jourdain, M. A. Valente, M. Brunet, J. Luterbacher, R. J. Allan, G. P. Compo, P. Jones, S. Brönnimann, and A. Lorrey acknowledge Package 3 of the Copernicus Climate Change Service 311a Lot1 for Collection and Processing of In Situ Observations Data Rescue. J. Luterbacher and R. J. Allan acknowledge Climate Science for Service Partnership China Project (CSSP). J. Luterbacher also acknowledges the DAAD project "The Mediterranean Hot-Spot" and the JPI-Climate/Belmont Forum collaborative Research Action 'INTEGRATE, An integrated data-model study of interactions between tropical monsoons and extratropical climate variability and extremes'. S. Brönnimann additionally acknowledges funding from the European Research Council under H2020 (grant PALAEO-RA, 787574).

GPCP data provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their web site at https://www.esrl.noaa.gov/psd/ (accessed 9 July 2019). The ERA-Interim, ERA-20C, and CERA-20C datasets are courtesy of ECMWF. The 20CR homepage is located at https://go.usa.gov/XTd. The scientific results and conclusions, as well as any views of opinions expressed herein, are those of the authors and do not necessarily reflect the views of the University of Colorado, NOAA, the Department of Commerce, or any other organization associated with this work.

ORCID

Laura C. Slivinski http://orcid.org/0000-0002-3531-3889
Gilbert P. Compo http://orcid.org/0000-0001-5199-9633
Linden Ashcroft http://orcid.org/0000-0003-3898-6648
Ed Hawkins http://orcid.org/0000-0001-9477-3677
Jürg Luterbacher http://orcid.org/0000-0002-8569-0973
G.W. Kent Moore http://orcid.org/0000-0002-3986-5605
Rajmund Przybylak http://orcid.org/0000-0003-4101-6116
Przemysław Wyszyński http://orcid.org/0000-0003-3470-7349

REFERENCES

Adler, R., Huffman, G., Chang, A., Ferraro, R., Xie, P., Janowiak, J., Rudolf, B., Schneider, U., Curtis, S., Bolvin, D., Gruber, A., Susskind, J. and Arkin, P. (2003) The version 2 Global Precipitation Climatology Project (GPCP) monthly precipitation analysis (1979–present). *Journal of Hydrometeorology*, 4, 1147–1167.

Allan, R.J. and Ansell, T. (2006) A new globally complete monthly historical gridded mean sea level pressure dataset (HadSLP2):

1850–2004. *Journal of Climate*, 19, 5816–5842. https://doi.org/10.1175/JCLI3937.1.

Allan, R.J., Brohan, P., Compo, G.P., Stone, R., Luterbacher, J. and Brönnimann, S. (2011) The International Atmospheric Circulation Reconstructions over the Earth (ACRE) initiative. *Bulletin of the American Meteorological Society*, 92, 1421–1425. https://doi.org/10.1175/2011BAMS3218.1.

Anderson, J.L. (2007) Exploring the need for localization in ensemble data assimilation using a hierarchical ensemble filter. *Physica D*, 230, 99–111.

Anderson, J.L. and Anderson, S.L. (1999) A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Monthly Weather Review*, 127, 2741–2758.

Anderson, E. and Järvinen, H. (1999) Variational quality control. *Quarterly Journal of the Royal Meteorological Society*, 125, 697–722.

Ansell, T.J., Jones, P.D., Allan, R.J., Lister, D., Parker, D.E., Brunet, M., Moberg, A., Jacobeit, J., Brohan, P., Rayner, N.A., Aguilar, E., Alexandersson, H., Barriendos, M., Brandsma, T., Cox, N.J., Della-Marta, P.M., Drebs, A., Founda, D., Gerstengarbe, F., Hickey, K., Jónsson, T., Luterbacher, J., Nordli, Ø., Oesterle, H., Petrakis, M., Philipp, A., Rodwell, M.J., Saladie, O., Sigro, J., Slonosky, V., Srnec, L., Swail, V., García-Suárez, A.M., Tuomenvirta, H., Wang, X., Wanner, H., Werner, P., Wheeler, D. and Xoplaki, E. (2006) Daily mean sea level pressure reconstructions for the European–North Atlantic region for the period 1850–2003. *Journal of Climate*, 19, 2717–2742. https://doi.org/10.1175/JCLI3775.1.

Bengtsson, L., Hagemann, S. and Hodges, K.I. (2004) Can climate trends be calculated from reanalysis data?. *Journal of Geophysical Research*, 109. https://doi.org/10.1029/2004JD004536.

Bloom, S.C., Takacs, L.L., da Silva, A.M. and Ledvina, D. (1996) Data assimilation using incremental analysis updates. *Monthly Weather Review*, 124, 1256–1271. https://doi.org/10.1175/1520-0493(1996)124<1256:DAUIAU>2.0.CO;2.

Bosilovich, M.G., Robertson, F.R. and Chen, J. (2011) Global energy and water budgets in MERRA. *Journal of Climate*, 24, 5721–5739. https://doi.org/10.1175/2011JCLI4175.1.

Brohan, P., Compo, G.P., Brönnimann, S., Allan, R.J., Auchmann, R., Brugnara, Y., Sardeshmukh, P.D. and Whitaker, J.S. (2016) The 1816 'year without a summer' in an atmospheric reanalysis. *Climate of the Past Discussions*, 2016, 1–11. https://doi.org/10.5194/cp-2016-78.

Brohan, P. (Personal communication, 22 March 2017).

Brönnimann, S., Brugnara, Y., Allan, R.J., Brunet, M., Compo, G.P., Crouthamel, R.I., Jones, P.D., Jourdain, S., Luterbacher, J., Siegmund, P., Valente, M.A. and Wilkinson, C.W. (2018) A roadmap to climate data rescue services. *Geoscience Data Journal*, 5, 28–39. https://doi.org/10.1002/gdj3.56.

Brönnimann, S., Grant, A.N., Compo, G.P., Ewen, T., Griesser, T., Fischer, A.M., Schraner, M. and Stickler, A. (2012) A multi-data set comparison of the vertical structure of temperature variability and change over the Arctic during the past 100 years. *Climate Dynamics*, 39, 1577–1598. https://doi.org/10.1007/s00382-012-1291-6.

Burn, M.J. and Palmer, S.E. (2015) Atlantic hurricane activity during the last millennium. *Scientific Reports*, 5. https://doi.org/10.1038/srep12838.

Brugnara, Y., Auchmann, R., BrÃ¶nnimann, S., Allan, R.J., Auer, I., Barriendos, M., Bergstrom, H., Bhend, J., Brázdil, R., Compo, G.P., Cornes, R.C., Dominguez-Castro, F., van Engelen, A.F.V., Filipiak, J., Holopainen, J., Jourdain, S., Kunz, M., Luterbacher, J., Maugeri, M., Mercalli, L., Moberg, A., Mock, C.J., Pichard, G., Řezníčková, L., van der Schrier, G., Slonosky, V., Ustrnul, Z., Valente, M.A., Wypych, A. and Yin, X. (2015) A collection of sub-daily pressure and temperature observations for the early instrumental periodwith a focus on the" year without a summer" 1816. *Climate of the Past*, 11, 1027–1047. https://doi.org/10.5194/cp-11-1027-2015.

Buehner, M. and Shlyaeva, A. (2015) Scale-dependent background-error covariance localisation. *Tellus A*, 67. https://doi.org/10.3402/tellusa.v67.28027.

Cionni, I., Eyring, V., Lamarque, J.F., Randel, W.J., Stevenson, D.S., Wu, F., Bodeker, G.E., Shepherd, T.G., Shindell, D.T. and Waugh, D.W. (2011) Ozone database in support of CMIP5 simulations: results and corresponding radiative forcing. *Atmospheric Chemistry and Physics*, 11, 11267–11292. https://doi.org/10.5194/acp-11-11267-2011.

Coddington, O., Lean, J.L., Pilewskie, P., Snow, M. and Lindholm, D. (2016) A solar irradiance climate data record. *Bulletin of the American Meteorological Society*, 97, 1265–1282.

Compo, G.P., Sardeshmukh, P.D. and Penland, C. (2001) Changes of subseasonal variability associated with El Niño. *Journal of Climate*, 14, 3356–3374.

Compo, G.P., Whitaker, J.S. and Sardeshmukh, P.D. (2006) Feasibility of a 100-year reanalysis using only surface pressure data. *Bulletin of the American Meteorological Society*, 87, 175–190.

Compo, G.P., Whitaker, J.S., Sardeshmukh, P.D., Matsui, N., Allan, R.J., Yin, X., Gleason, B.E., Vose, R.S., Rutledge, G., Bessemoulin, P., Brönnimann, S., Brunet, M., Crouthamel, R.I., Grant, A.N., Groisman, P.Y., Jones, P.D., Kruk, M.C., Kruger, A.C., Marshall, G.J., Maugeri, M., Mok, H.Y., Nordli, Ø., Ross, T.F., Trigo, R.M., Wang, X.L., Woodruff, S.D. and Worley, S.J. (2011) The Twentieth Century Reanalysis project. *Quarterly Journal of the Royal Meteorological Society*, 137, 1–28.

Compo, G.P., Sardeshmukh, P.D., Whitaker, J.S., Brohan, P., Jones, P.D. and McColl, C. (2013) Independent confirmation of global land warming without the use of station temperatures. *Geophysical Research Letters*, 40, 3170–3174. https://doi.org/10.1002/grl.50425.

Compo, G.P., Whitaker, J.S., Sardeshmukh, P.D., Allan, R.J., McColl, C., Yin, X., Vose, R.S., Matsui, N., Ashcroft, L., Auchmann, R., Benoy, M., Bessemoulin, P., Brandsma, T., Brohan, P., Brunet, M., Comeaux, J., Cram, T.A., Crouthamel, R., Groisman, P.Y., Hersbach, H., Jones, P.D., Jonsson, T., Jourdain, S., Kelly, G., Knapp, K.R., Kruger, A., Kubota, H., Lentini, G., Lorrey, A., Lott, N., Lubker, S.J., Luterbacher, J., Marshall, G.J., Maugeri, M., Mock, C.J., Mok, H.Y., Nordli, O., Przybylak, R., Rodwell, M.J., Ross, T.F., Schuster, D., Srnec, L., Valente, M.A., Vizi, Z., Wang, X.L., Westcott, N., Woollen, J.S. and Worley, S.J. (2015) The International Surface Pressure Databank Version 3. NCAR/UCAR Research Data Archive. https://doi.org/10.5065/D6D50K29; accessed 22 July 2019.

Cram, T.A., Compo, G.P., Yin, X., Allan, R.J., McColl, C., Vose, R.S., Whitaker, J.S., Matsui, N., Ashcroft, L., Auchmann, R., Bessemoulin, P., Brandsma, T., Brohan, P., Brunet, M., Comeaux, J., Crouthamel, R., Gleason, B.E., Groisman, P.Y., Hersbach, H., Jones, P.D., Jónsson, T., Jourdain, S., Kelly, G., Knapp, K.R., Kruger, A., Kubota, H., Lentini, G., Lorrey, A., Lott, N., Lubker, S.J., Luterbacher, J., Marshall, G.J., Maugeri, M., Mock, C.J., Mok, H.Y., Nordli, Ø., Rodwell, M.J., Ross, T.F., Schuster, D., Srnec, L., Valente, M.A., Vizi, Z., Wang, X.L., Westcott, N., Woollen, J.S. and Worley, S.J. (2015) The international surface pressure databank version 2. *Geoscience Data Journal*, 2, 31–46.

Crowley, T.J. and Unterman, M.B. (2012) Technical details concerning development of a 1200-yr proxy index for global volcanism. *Earth*

System Science Data, 5, 187–197. https://doi.org/10.5194/essdd-5-1-2012.

Daley, R. (1993) *Atmospheric Data Analysis*. Cambridge University Press, Cambridge, UK.

Dee, D.P., Uppala, S., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A.C.M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A.J., Haimberger, L., Healy, S.B., Hersbach, H., Hólm, E.V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A.P., Monge-Sanz, B.M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N. and Vitart, F. (2011) The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137, 553–597.

Deser, C., Simpson, I.R., McKinnon, K.A. and Phillips, A.S. (2017) The Northern Hemisphere extratropical atmospheric circulation response to ENSO: how well do we know it and how do we evaluate models accordingly?. *Journal of Climate*, 30, 5059–5082.

Desroziers, G., Berre, L., Chapnik, B. and Poli, P. (2005) Diagnosis of observation, background and analysis error statistics in observation space. *Quarterly Journal of the Royal Meteorological Society*, 131, 3385–3396. https://doi.org/10.1256/qj.05.108.

Dharssi, I., Lorenc, A.C. and Ingleby, N.B. (1992) Treatment of gross errors using maximum probability theory. *Quarterly Journal of the Royal Meteorological Society*, 118, 1017–1036.

Diamond, H., Lorrey, A., Knapp, K. and Levinson, D. (2012) Development of an enhanced tropical cyclone tracks database for the Southwest Pacific from 1840 to 2010. *International Journal of Climatology*, 32, 2240–2250.

Diaz, H., Folland, C.K., Manabe, T., Parker, D.E., Reynolds, R. and Woodruff, S.D. (2002) Workshop on advances in the use of historical marine climate data. *Bulletin of the World Meteorological Organization*, 51, 377–379.

Donat, M.G., Alexander, L.V., Herold, N. and Dittus, A.J. (2016) Temperature and precipitation extremes in century-long gridded observations, reanalyses, and atmospheric model simulations. *Journal of Geophysical Research: Atmospheres*, 121, 11174–11189. https://doi.org/10.1002/2016JD025480.

Ek, M., Mitchell, K., Lin, Y., Rogers, E., Grunmann, P., Koren, V., Gayno, G. and Tarpley, J. (2003) Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. *Journal of Geophysical Research: Atmospheres*, 108(D22). https://doi.org/10.1029/2002JD003296.

Ferguson, C.R. and Villarini, G. (2012) Detecting inhomogeneities in the twentieth century reanalysis over the central United States. *Journal of Geophysical Research*, 117. https://doi.org/10.1029/2011JD016988.

Freeman, E., Woodruff, S.D., Worley, S.J., Lubker, S.J., Kent, E.C., Angel, W.E., Berry, D.I., Brohan, P., Eastman, R., Gates, L., Gloeden, W., Ji, Z., Lawrimore, J., Rayner, N.A., Rosenhagen, G. and Smith, S.R. (2017) ICOADS Release 3.0: a major update to the historical marine climate record. *International Journal of Climatology*, 37, 2211–2232. https://doi.org/10.1002/joc.4775.

Fujiwara, M., Wright, J.S., Manney, G.L., Gray, L.J., Anstey, J., Birner, T., Davis, S., Gerber, E.P., Harvey, V.L., Hegglin, M.I., Homeyer, C.R., Knox, J.A., Krüger, K., Lambert, A., Long, C.S., Martineau, P., Molod, A., Monge-Sanz, B.M., Santee, M.L., Tegtmeier, S., Chabrillat, S., Tan, D.G.H., Jackson, D.R., Polavarapu, S., Compo, G.P., Dragani, R., Ebisuzaki, W., Harada, Y., Kobayashi, C., McCarty, W., Onogi, K., Pawson, S., Simmons, A.J., Wargan, K., Whitaker, J.S. and Zou, C.-Z. (2017) Introduction to the SPARC Reanalysis Intercomparison Project (S-RIP) and overview of the reanalysis systems. *Atmospheric Chemistry and Physics*, 17, 1417–1452.

Gaspari, G. and Cohn, S.E. (1999) Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, 125, 723–757.

Gelaro, R., McCarty, W., Suárez, M.J., Todling, R., Molod, A., Takacs, L., Randles, C.A., Darmenov, A., Bosilovich, M.G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., da Silva, A.M., Gu, W., Kim, G.-K., Koster, R., Lucchesi, R., Merkova, D., Nielsen, J.E., Partyka, G., Pawson, S., Putman, W., Rienecker, M., Schubert, S.D., Sienkiewicz, M. and Zhao, B. (2017) The Modern-Era Retrospective analysis for Research and Applications, version 2 (MERRA-2). *Journal of Climate*, 30, 5419–5454.

Giese, B.S., Compo, G.P., Slowey, N.C., Sardeshmukh, P.D., Carton, J.A., Ray, S. and Whitaker, J.S. (2010) The 1918/19 El Niño. *Bulletin of the American Meteorological Society*, 91, 177–183.

Giese, B.S., Seidel, H.F., Compo, G.P. and Sardeshmukh, P.D. (2016) An ensemble of ocean reanalyses for 1815–2013 with sparse observational input. *Journal of Geophysical Research*, 121, 6891–6910. https://doi.org/10.1002/2016JC012079.

Häkkinen, S., Rhines, P.B. and Worthen, D.L. (2011) Atmospheric blocking and Atlantic multidecadal ocean variability. *Science*, 334, 655–659.

Hamill, T.M., Whitaker, J.S. and Snyder, C. (2001) Distance-dependent filtering of background-error covariance estimates in an ensemble Kalman filter. *Monthly Weather Review*, 129, 2776–2790.

Han, J., Witek, M.L., Teixeira, J., Sun, R., Pan, H.-L., Fletcher, J.K. and Bretherton, C.S. (2016) Implementation in the NCEP GFS of a hybrid eddy-diffusivitymass-flux (EDMF) boundary-layer parameterization with dissipative heating and modified stable boundary-layer mixing. *Weather and Forecasting*, 31, 341–352.

Harris, L.M. and Lin, S.-J. (2013) A two-way nested global-regional dynamical core on the cubed-sphere grid. *Monthly Weather Review*, 141, 283–306.

Harvey, T., Renwick, J.A., Lorrey, A.M. and Ngari, A. (2019) The representation of the South Pacific convergence zone in the Twentieth Century Reanalysis. *Monthly Weather Review*, 147, 841–851.

Hegerl, G.C., Brönnimann, S., Schurer, A. and Cowan, T. (2018) The early 20th century warming: anomalies, causes, and consequences. *Wiley Interdisciplinary Reviews*, 9(4). https://doi.org/10.1002/wcc.522.

Hirahara, S., Ishii, M. and Fukuda, Y. (2014) Centennial-scale sea surface temperature analysis and its uncertainty. *Journal of Climate*, 27, 57–75.

Hou, Y.-T., Moorthi, S. and Compana, K. (2002) Parameterization of solar radiation transfer in the NCEP models. Office note 441, NCEP, Washington DC. http://www.lib.ncep.noaa.gov/ncepofficenotes/2000s; accessed 22 July 2019

Houtekamer, P.L. and Mitchell, H.L. (1998) Data assimilation using an ensemble Kalman filter technique. *Monthly Weather Review*, 126, 796–811.

Houtekamer, P.L. and Mitchell, H.L. (2001) A sequential ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review*, 129, 123–137.

Huang, J., Ji, M., Xie, Y., Wang, S., He, Y. and Ran, J. (2016) Global semi-arid climate change over last 60 years. *Climate Dynamics*, 46, 1131–1150. https://doi.org/10.1007/s00382-015-2636-8.

Huang, J., Xie, Y., Guan, X., Li, D. and Ji, F. (2017) The dynamics of the warming hiatus over the Northern Hemisphere. *Climate Dynamics*, 48, 429–446. https://doi.org/10.1007/s00382-016-3085-8.

Huang, X.-Y. and Lynch, P. (1993) Diabatic digital-filtering initialization: application to the HIRLAM model. *Monthly Weather Review*, 121, 589–603. https://doi.org/10.1175/1520-0493(1993)121<0589:DDFIAT>2.0.CO;2.

Janjić, T. and Cohn, S.E. (2006) Treatment of observation error due to unresolved scales in atmospheric data assimilation. *Monthly Weather Review*, 134, 2900–2915.

Jones, P.D., Harpham, C. and Briffa, K. (2013) Lamb weather types derived from reanalysis products. *International Journal of Climatology*, 33, 1129–1139.

Jones, P.D., Harpham, C. and Lister, D. (2016) Long-term trends in gale days and storminess for the Falkland islands. *International Journal of Climatology*, 36, 1413–1427.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K.C., Ropelewski, C., Wang, J., Leetma, A., Reynolds, R., Jenne, R. and Joseph, D. (1996) The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, 77, 437–472.

Kanamitsu, I., Alpert, J., Campana, K., Caplan, P., Deaven, D., Iredell, M., Katz, B., Pan, H.-L., Sela, J. and White, G. (1991) Recent changes implemented into the global forecast system at NMC. *Weather and Forecasting*, 6, 425–435.

Kent, E.C., Fangohr, S. and Berry, D.I. (2013) A comparative assessment of monthly mean wind speed products over the global ocean. *International Journal of Climatology*, 33, 2520–2541. https://doi.org/10.1002/joc.3606.

Kinter, J., Fennessy, M., Krishnamurthy, V. and Marx, L. (2004) An evaluation of the apparent interdecadal shift in the tropical divergent circulation in the NCEP–NCAR reanalysis. *Journal of Climate*, 17, 349–361.

Kistler, R., Kalnay, E., Collins, W., Saha, S., White, G., Woollen, J., Chelliah, M., Ebisuzaki, W., Kanamitsu, M., Kousky, V., van den Dool, H., Jenne, R. and Fiorino, M. (2001) The NCEP–NCAR 50-year reanalysis: monthly means CD-ROM and documentation. *Bulletin of the American Meteorological society*, 82, 247–268.

Kleist, D.T. and Ide, K. (2015) An OSSE-based evaluation of hybrid variationalâŁ“ensemble data assimilation for the NCEP GFS. Part II: 4DEnVar and hybrid variants. *Monthly Weather Review*, 143, 452–470.

Klotzbach, P.J., Oliver, E.C.J., Leeper, R.D. and Schreck, C.J. (2016) The relationship between the Madden–Julian Oscillation (MJO) and southeastern New England snowfall. *Monthly Weather Review*, 144, 1355–1362. https://doi.org/10.1175/MWR-D-15-0434.1.

Knapp, K.R., Kruk, M.C., Levinson, D.H., Diamond, H.J. and Neumann, C.J. (2010) The international best track archive for climate stewardship (IBTrACS) unifying tropical cyclone data. *Bulletin of the American Meteorological Society*, 91, 363–376.

Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., Onogi, K., Kamahori, H., Kobayashi, C., Endo, H., Miyaoka, K. and Takahashi, K. (2015) The JRA-55 reanalysis: general specifications and basic characteristics. *Journal of the Meteorological Society of Japan. Series II*, 93, 5–48.

Kruk, M.C., Knapp, K.R. and Levinson, D.H. (2010) A technique for combining global tropical cyclone best-track data. *Journal of Atmospheric and Oceanic Technology*, 27, 680–692.

Kubota, H. (2012) Variability of typhoon tracks and genesis over the Western North Pacific, in *Cyclones, Formation, Triggers and Control*, Oouchi, K, Fudeyasu, H. (eds), pp. 95–114. Nova Science Publishers, Hauppauge, NY, USA.

Kwon, J., Yi, J.-W. and Song, S.M. (2004) Adaptive cubic interpolation of CT slices for maximum intensity projections, in *Medical Imaging 2004: Visualization, Image-Guided Procedures, and Display*, pp. 837–845. Galloway, R.L. (ed.), SPIE Proceedings volume 5367, Bellingham WA.

Laloyaux, P., de Boisseson, E., Balmaseda, M., Bidlot, J.-R., Br^{ſ}onnimann, S., Buizza, R., Dalhgren, P., Dee, D.P., Haimberger, L., Hersbach, H., Kosaka, Y., Martin, M., Poli, P., Rayner, N.A., Rustemeier, E. and Schepers, D. (2018) CERA-20C: a coupled reanalysis of the Twentieth Century. *Journal of Advances in Modeling Earth Systems*, 10, 1172–1195. https://doi.org/10.1029/2018MS001273.

Landschützer, P., Gruber, N., Haumann, F.A., Rödenbeck, C., Bakker, D.C.E., van Heuven, S., Hoppema, M., Metzl, N., Sweeney, C., Takahashi, T., Tilbrook, B. and Wanninkhof, R. (2015) The reinvigoration of the Southern Ocean carbon sink. *Science*, 349, 1221–1224.

Landsea, C. (Personal communication, 30th September 2008).

Landsea, C.W., Glenn, D.A., Bredemeyer, W., Chenoweth, M., Ellis, R., Gamache, J., Hufstetler, L., Mock, C., Perez, R., Prieto, R., Sanchez-Sesma, J., Thomas, D. and Woolcock, L. (2008) A reanalysis of the 1911-20 Atlantic hurricane database. *Journal of Climate*, 21, 2138–2168.

Lei, L. and Whitaker, J.S. (2016) A four-dimensional incremental analysis update for the ensemble Kalman filter. *Monthly Weather Review*, 144, 2605–2621.

Lorenc, A.C. (1986) Analysis methods for numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 112, 1177–1194. https://doi.org/10.1002/qj.49711247414.

Lorrey, A., Dalu, G., Renwick, J., Diamond, H. and Gaetani, M. (2012) Reconstructing the South Pacific convergence zone position during the presatellite era: a La Niña case study. *Monthly Weather Review*, 140, 3653–3668.

Lorrey, A.M. and Chappell, P.R. (2016) The “dirty weather” diaries of Reverend Richard Davis: insights about early colonial-era meteorology and climate variability for northern New Zealand, 1839–1851. *Climate of the Past*, 12, 553–553.

Lott, J.N., Vose, R.S., Del Greco, S.A., Ross, T.F., Worley, S. and Comeaux, J.L. (2008) The integrated surface database: partnerships and progress. In: *Extended Abstracts, 24th Conference on Interactive Information and Processing Systems*, 20–24 January 2008, New Orleans, LA. American Meteorological Society, Boston, MA.

Lynch, P. and Huang, X.-Y. (1992) Initialization of the HIRLAM model using a digital filter. *Monthly Weather Review*, 120, 1019–1034. https://doi.org/10.1175/1520-0493(1992)120<1019:IOTHMU>2.0.CO;2.

McCormack, J.P., Eckermann, S.D., Siskind, D.E. and McGee, T.J. (2006) CHEM2D-OPP: A new linearized gas-phased ozone photochemistry parameterization for high-altitude NWP and climate models. *Atmospheric Chemistry and Physics*, 6, 4943–4972. https://doi.org/10.5194/acp-6-4943-2006.

Mlawer, E.J., Taubman, S.J., Brown, P.D., Iacono, M.J. and Clough, S.A. (1997) Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave, 102, 16663–16682.

Middleton, W.E.K. (1964) *The History of the Barometer*. Johns Hopkins University Press, Baltimore, MD.

Monmonier, M. (1999) *Air Apparent: How Meteorologists Learned to Map, Predict, and Dramatize Weather*. University of Chicago Press, Chicago, IL.

Moore, G. and Babij, M. (2017) Iceland's great frost winter of 1917/1918 and its representation in reanalyses of the Twentieth Century. *Quarterly Journal of the Royal Meteorological Society*, 143, 508–520.

Moorthi, S. (Personal communication, 31st March 2011).

Moorthi, S., Pan, H.-L. and Caplan, P. (2001) Changes to the 2001 NCEP operational MRF/AVN global analysis/forecast system. NWS Technical Procedures Bulletin 484, NCEP, Washington DC.

Morice, C.P., Kennedy, J.J., Rayner, N.A. and Jones, P.D. (2012) Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: the HADCRUT4 data set. *Journal of Geophysical Research: Atmospheres*, 117. https://doi.org/10.1029/2011JD017187.

Mueller, W.A., Pohlmann, H., Sienz, F. and Smith, D. (2014) Decadal climate predictions for the period 1901–2010 with a coupled climate model. *Geophysical Research Letters*, 41, 2100–2107.

Paik, S. and Min, S.-K. (2017) Climate responses to volcanic eruptions assessed from observations and CMIP5 multi-models. *Climate Dynamics*, 48, 1017–1030. https://doi.org/10.1007/s00382-016-3125-4.

Parker, W.S. (2016) Reanalyses and observations: what's the difference?. *Bulletin of the American Meteorological Society*, 97, 1565–1572.

Palmer, T.N., Buizza, R., Doblas-Reyes, F., Jung, T., Leutbecher, M., Shutts, G.J., Steinheimer, M. and Weisheimer, A. (2009) Stochastic parametrization and model uncertainty. Technical Memorandum 598, ECMWF, Reading, UK.

Pegion, P. (Personal communication, 2nd February 2017).

Poli, P., Hersbach, H., Berrisford, P., Dee, D.P., Simmons, A.J. and Laloyaux, P. (2015) ERA-20C Deterministic. ERA Report 20, ECMWF, Reading, UK.

Poli, P., Hersbach, H., Dee, D.P., Berrisford, P., Simmons, A.J., Vitart, F., Laloyaux, P., Tan, D.G., Peubey, C., Thépaut, J.-N., Trémolet, Y., Hólm, E.V., Bonavita, M., Isaksen, L. and Fisher, M. (2016) ERA-20C: an atmospheric reanalysis of the Twentieth Century. *Journal of Climate*, 29, 4083–4097.

Reynolds, R.W., Smith, T.M., Liu, C., Chelton, D.B., Casey, K.S. and Schlax, M.G. (2007) Daily high-resolution-blended analyses for sea surface temperature. *Journal of Climate*, 20, 5473–5496.

Saha, S., Moorthi, S., Pan, H.-L., Wu, X., Wang, J., Nadiga, S., Tripp, P., Kistler, R., Woollen, J., Behringer, D., Liu, H., Stokes, D., Grumbine, R., Gayno, G., Hou, Y.-T., Chuang, H.-Y., Juang, H.-M.H., Sela, J., Iredell, M., Treadon, R., Kleist, D., van Delst, P., Keyser, D., Derber, J., Ek, M., Meng, J., Wei, H., Yang, R., Lord, S., van den Dool, H., Kumar, A., Wang, W., Long, C., Chelliah, M., Xue, Y., Huang, B., Schemm, J.-K., Ebisuzaki, W., Lin, R., Xie, P., Chen, M., Zhou, S., Higgins, W., Zou, C.-Z., Liu, Q., Chen, Y., Han, Y., Cucurull, L., Reynolds, R.W., Rutledge, G. and Goldberg, M. (2010) The NCEP climate forecast system reanalysis. *Bulletin of the American Meteorological Society*, 91, 1015–1058.

Shutts, G.J., Leutbecher, M., Weisheimer, A., Stockdale, T., Isaksen, L. and Bonavita, M. (2011) Representing model uncertainty: stochastic parameterizations at ECMWF. *ECMWF Newsletter*, 129, 19–24.

Raphael, M.N. (2004) A zonal wave 3 index for the Southern Hemisphere. *Geophysical Research Letters*, 31(23). https://doi.org/10.1029/2004GL020365.

Rayner, N.A., Brohan, P., Parker, D.E., Folland, C.K., Kennedy, J.J., Vanicek, M., Ansell, T.J. and Tett, S.F.B. (2006) *Improved analyses of changes and uncertainties in sea surface temperature measured in situ since the mid-nineteenth century: the HadSST2 dataset*, Vol. 19 pp. 446–469. https://doi.org/10.1175/JCLI3637.1.

Rohrer, M., Brönnimann, S., Martius, O., Raible, C.C., Wild, M. and Compo, G.P. (2018) Representation of extratropical cyclones, blocking anticyclones, and alpine circulation types in multiple reanalyses and model simulations. *Journal of Climate*, 31, 3009–3031. https://doi.org/10.1175/JCLI-D-17-0350.1.

Slivinski, L.C. (2018) Historical reanalysis: what, how, and why?. *Journal of Advances in Modeling Earth Systems*, 10, 1736–1739. https://doi.org/10.1029/2018MS001434.

Slonosky, V.C. and Graham, E. (2005) Canadian pressure observations and circulation variability: links to air temperature. *International Journal of Climatology*, 25, 1473–1492.

Smith, A., Lott, N. and Vose, R. (2011) The integrated surface database: recent developments and partnerships. *Bulletin of the American Meteorological Society*, 92, 704–708.

Spencer, L., McColl, C., Brohan, P., Wood, K., Allan, R.J. and Compo, G.P. (2019) OldWeather3 Marine Data for ICOADS Input. Digital Asset Services Hub, NCAR/UCAR, Boulder, CO.

Tavolato, C. and Isaksen, L. (2015) On the use of a Huber norm for observation quality control in the ECMWF 4D-Var. *Quarterly Journal of the Royal Meteorological Society*, 141, 1514–1527.

Titchner, H.A. and Rayner, N.A. (2014) The Met Office Hadley Centre sea ice and sea surface temperature data set, version 2: 1. Sea ice concentrations. *Journal of Geophysical Research*, 119, 2864–2889. https://doi.org/10.1002/2013JD020316.

Tompkins, A. and Berner, J. (2008) A stochastic convective approach to account for model uncertainty due to unresolved humidity variability. *Journal of Geophysical Research: Atmospheres*, 113(D18). https://doi.org/10.1029/2007JD009284.

Trenberth, K.E. and Paolino, D.A. (1980) The Northern Hemisphere sea-level pressure data set: trends, errors and discontinuities. *Monthly Weather Review*, 108, 855–872. https://doi.org/10.1175/1520-0493(1980)108<0855:TNHSLP>2.0.CO;2.

Trenberth, K.E. and Smith, L. (2005) The mass of the atmosphere: a constraint on global analyses. *Journal of Climate*, 18, 864–875.

Ulbrich, U., Fink, A.H., Klawa, M. and Pinto, J.G. (2001) Three extreme storms over Europe in December 1999. *Weather*, 56, 70–80.

van den Besselaar, E.J.M., Haylock, M.R., van der Schrier, G. and Klein Tank, A.M.G. (2011) A European daily high-resolution observational gridded data set of sea level pressure. *Journal of Geophysical Research: Atmospheres*, 116(D11). https://doi.org/10.1029/2010JD015468.

van Loon, H. and Jenne, R.L. (1972) The zonal harmonic standing waves in the Southern Hemisphere. *Journal of Geophysical Research*, 77, 992–1003. https://doi.org/10.1029/JC077i006p00992.

Walsh, J.E., Chapman, W.L. and Fetterer, F. (2016) Gridded Monthly Sea Ice Extent and Concentration, 1850 Onward, Version 1. National Snow and Ice Data Center, Boulder, CO. https://nsidc.org/data/g10010; accessed 7 July 2019

Wang, J.-W.A., Sardeshmukh, P.D., Compo, G.P., Whitaker, J.S., Slivinski, L.C., McColl, C.M. and Pegion, P.J. (2019) Sensitivities of the NCEP Global Forecast System, (Vol. 147

Wang, X.L., Feng, Y., Chan, R. and Isaac, V. (2016) Intercomparison of extratropical cyclone activity in nine reanalysis datasets. *Atmospheric Research*, 181, 133–153.

Wang, X.L., Feng, Y., Compo, G.P., Swail, V.R., Zwiers, F.W., Allan, R.J. and Sardeshmukh, P.D. (2013) Trends and low-frequency variability of extratropical cyclone activity in the ensemble of Twentieth Century reanalysis. *Climate Dynamics*, 40, 2775–2800.

Whitaker, J.S., Compo, G.P., Wei, X. and Hamill, T.M. (2004) Reanalysis without radiosondes using ensemble data assimilation. *Monthly Weather Review*, 132, 1190–1200.

Whitaker, J.S. and Hamill, T.M. (2002) Ensemble data assimilation without perturbed observations. *Monthly Weather Review*, 130, 1913–1924.

Whitaker, J.S. and Hamill, T.M. (2012) Evaluating methods to account for system errors in ensemble data assimilation. *Monthly Weather Review*, 140, 3078–3089.

Winton, M. (2000) A reformulated three-layer sea ice model. *Journal of Atmospheric and Oceanic Technology*, 17, 525–531.

Woodruff, S.D., Worley, S.J., Lubker, S.J., Ji, Z., Freeman, J.E., Berry, D.I., Brohan, P., Kent, E.C., Reynolds, R.W., Smith, S.R. and Wilkinson, C. (2011) ICOADS Release 2.5: extensions and enhancements to the surface marine meteorological archive. *International Journal of Climatology*, 31, 951–967. https://doi.org/10.1002/joc.2103.

Worley, S.J., Woodruff, S.D., Reynolds, R.W., Lubker, S.J. and Lott, N. (2005) ICOADS release 2.1 data and products. *International Journal of Climatology*, 25, 823–842. https://doi.org/10.1002/joc.1166.

Yin, X., Gleason, B.E., Compo, G.P., Matsui, N. and Vose, R.S. (2008) The International Surface Pressure Databank (ISPD) land component. Instruction manual version 2.2. National Climatic Data Center: Asheville, NC.

Zhang, L., Kumar, A. and Wang, W. (2012) Influence of changes in observations on precipitation: a case study for the Climate Forecast System Reanalysis (CFSR). *Journal of Geophysical Research*, 117. https://doi.org/10.1029/2011JD017347.

How to cite this article: Slivinski LC, Compo GP, Whitaker JS, et al. Towards a more reliable historical reanalysis: Improvementsfor version 3 of the Twentieth Century Reanalysis system. *Q J R Meteorol Soc*. 2019;145:2876–2908. https://doi.org/10.1002/qj.3598

APPENDICES

A.1 | Implementation details for 20CR versions 2c and 3

As several of the unique details of the 20CRv2c system have not been previously documented, they are described here in detail by way of comparison with the new 20CRv3 system. 20CRv2c used an experimental version of the 2008 NCEP GFS (2008EX), a spectral coupled atmosphere–land model with a comprehensive suite of physical parametrizations and processes including representing the radiative effects of time-varying CO_2 concentrations, volcanic aerosols, and an 11-year solar cycle (Compo *et al.*, 2011), at a horizontal resolution of total spherical wavenumber 62 (about 2°) with 28 vertical hybrid sigma-pressure levels. In this version of the GFS, the NCEP Global Spectral Model atmosphere is coupled to the Noah land surface model (Ek *et al.*, 2003) with four subsurface soil levels and to a 2.5-layer thermodynamic ice model with two layers of ice and one layer of snow (Winton, 2000). The physical parametrizations used in the GFS 2008EX are described by Kanamitsu *et al.* (1991) and Moorthi *et al.* (2001). Additional updates to these parametrizations, including revised solar radiative transfer, boundary-layer vertical diffusion, cumulus convection, and gravity wave drag parametrizations, are described by Saha *et al.* (2010). In addition, the cloud liquid water is a prognostic quantity with a simple cloud microphysics parametrization. The radiation interacts with a fractional cloud cover that is diagnostically determined by the predicted cloud liquid water. Radiation also interacts with a three-dimensional prognostic ozone that is determined from a parametrization of gas-phase ozone production and loss chemistry (McCormack *et al.*, 2006) implemented by NCEP/EMC (Moorthi, personal communication, 2011) The GFS 2008EX also includes the radiative effects of historical time-varying CO_2 concentrations, volcanic aerosols, and solar variations using the long-wave radiation model of Mlawer *et al.* (1997) and short-wave radiation model of Hou *et al.* (2002).

The specified boundary conditions needed to run the coupled atmosphere–land model are taken from time-evolving sea surface temperature and sea ice concentration fields. The sea ice concentration fields are taken from monthly COBE-SST2 (Hirahara *et al.*, 2014) interpolated to daily values. The GFS 2008EX was modified to allow sea ice concentration specification down to fractions of 0.15, where previously the smallest concentration allowed was 0.5. Additionally, the sea ice concentration mis-specification noted by Compo *et al.* (2011) was corrected. The SST fields for 1851–2012 consist of 18 members of pentad Simple Ocean Data Assimilation with sparse input version 2 (SODAsi.2; Giese *et al.*, 2016) with the high latitudes (above and below 60°) adjusted to COBE-SST2 (Hirahara *et al.*, 2014), and the 1981–2010 climatology adjusted at each gridpoint to that of COBE-SST2. The pentad values were interpolated to daily resolution. Years 2013–2014 have specified SST fields from the Optimum Interpolation SST version 2 (OISSTv2; Reynolds *et al.*, 2007); the reanalysis was not extended beyond 2014.

20CRv2c used additional pressure observations and modified observation errors compared to those used in 20CRv2 and described by Compo *et al.* (2011). Observations are from the International Surface Pressure Databank (ISPD) version 3.2.9 (Cram *et al.*, 2015; Compo *et al.*, 2015), which consists of station observations, marine observations, and pressure reports for tropical cyclones from the International Best Track Archive for Climate Stewardship (IBTrACS) v03r05 (Knapp *et al.*, 2010; Kruk *et al.*, 2010). The station component is a blend of many national and international collections, with the largest contributor being surface and SLP observations from the Integrated Surface Database (ISD; Lott *et al.* 2008; Smith *et al.* 2011). Procedures for blending the station component

TABLE A1 Observation errors (hPa) assigned to tropical cyclone data in 20CRv2c. The upper half of the table is reserved for observed central pressure data, and the lower half of the table includes wind-derived central pressure data. Columns denote basin and storm category, from 'tropical depression' (TD) to category 5.

	Basin	TD	Cat1	Cat2	Cat3	Cat4	Cat5
Observed	b1	6.0	6.0	7.0	9.0	9.0	9.0
	b2	6.0	6.0	7.0	9.0	9.0	9.0
	b3	3.0	3.0	3.5	4.5	4.5	4.5
	b4	3.0	3.0	3.5	4.5	4.5	4.5
	b5	9.0	9.0	10.5	13.5	13.5	13.5
	b6	9.0	9.0	10.5	13.5	13.5	13.5
	b7	9.0	9.0	10.5	13.5	13.5	13.5
	b8	6.0	6.0	7.0	9.0	9.0	9.0
	b9	3.0	3.0	3.5	4.5	4.5	4.5
Wind-	b1	6.0	6.5	7.0	9.0	9.0	9.0
derived	b2	6.0	6.5	7.0	9.0	9.0	9.0
	b3	4.0	5.0	5.5	6.0	6.0	6.0
	b4	4.0	5.0	5.5	6.0	6.0	6.0
	b5	12.0	13.0	14.0	18.0	18.0	18.0
	b6	12.0	13.0	14.0	18.0	18.0	18.0
	b7	12.0	13.0	14.0	18.0	18.0	18.0
	b8	6.0	6.5	7.0	9.0	9.0	9.0
	b9	4.0	5.0	5.5	6.0	6.0	6.0

b1. Eastern North Pacific (sub-basin: Central Pacific).
b2. Eastern North Pacific.
b3. North Atlantic.
b4. North Indian.
b5. South Indian
b6. South Pacific (sub-basin: Eastern Australia).
b7. South Pacific.
b8. West Pacific.
b9. South Indian (sub-basin: Western Australia)

are described by Yin *et al.* (2008). Marine observations are from the International Comprehensive Ocean–Atmosphere Data Set (ICOADS; Worley *et al.* 2005; Woodruff *et al.* 2011) version 2.5.P, blended with additional observations from ships of exploration during the mid-19th to mid-20th centuries (Allan *et al.*, 2011; Cram *et al.*, 2015). The surface pressure observations have been made available through international cooperation facilitated by the Atmospheric Circulation Reconstructions over the Earth (ACRE) initiative and working groups of the Global Climate Observing System and World Climate Research Programme. Many observations from the early 19th century were collected as part of a study on the Tambora eruption of 1815 (Brugnara *et al.*, 2015; Brohan *et al.*, 2016). The observation errors used in 20CRv2c are the same as described in Table 1, with the exception of tropical cyclones; the errors used in 20CRv2c for those data were defined based on storm category, ocean basin/sub-basin, and whether the assimilated value was an observed central pressure or a wind-derived central pressure (Landsea *et al.*, 2008; Compo *et al.*, 2011; Kubota, 2012 and Landsea, personal communication, 2008). These errors are given in Table A.1.

As in Compo *et al.* (2011), every fifth year was produced in parallel for a continuous 5-year stream that started from November of (stream year –1). Stream years are 1850, 1855, ..., 2000, 2005. Released data start in January of stream year +1, e.g. 1851 to 1855, 1856 to 1860, ..., 2001 to 2005. Stream year 2005 is extended to 2012 with the SODAsi.2 SST fields and then to 2014 with the OISSTv2 SST fields.

While still assimilating only pressure observations, the NOAA-CIRES-DOE 20CRv3 has changes in all other aspects of the DA compared to 20CRv2c. 20CRv3 uses an updated version of the NCEP GFS, v14.0.1, including the same Noah land surface model and thermodynamic ice model used in used in 2008EX and many other previous versions of the GFS. Version 14.0.1 of the GFS became operational at NCEP in autumn 2017. Several adjustments were made to the operational model prior to implementation in the 20CRv3 system. First, the operational model includes an ensemble run at a resolution of T574, with a single deterministic forecast run at a resolution of T1534. The version of the GFS used in 20CRv3 is run at a resolution of T254. Second, the dry air mass is specified to be 98.305 kPa (Trenberth and Smith, 2005). Third, as in 20CRv2c, sea ice concentrations are allowed down to 0.15. Fourth, the radiation interacts with CMIP5 ozone from 1850 onwards (Cionni *et al.*, 2011); prior to 1850, it uses 1850-level CMIP5 ozone. The model still advances a prognostic ozone determined from a gas-phase parametrization of linearized ozone production and loss (McCormack *et al.*, 2006) implemented by NCEP/EMC (Moorthi, personal communication, 2011) This prognostic ozone is output during model forecasts, but is not used in the internal radiation computations. This was done to prevent spurious trends associated with the fact that the prognostic ozone scheme was developed for conditions that existed in the late 20th century, including ozone depletion and the 'ozone hole' associated with CFC emissions. Next, the solar forcing in 20CRv3 is determined from the Total Solar Irradiance Reconstruction based on NRLTSI2 (Coddington *et al.*, 2016). Volcanic aerosols in 20CRv3 are prescribed according to Crowley and Unterman (2012). A hybrid eddy-diffusivity mass-flux boundary-layer parametrization was used as described by Han *et al.* (2016), but the dissipative heating in tropical cyclones that is used operationally was turned off for 20CRv3 (Pegion, personal communication, 2017) The coefficients that determine the auto-conversion from ice to snow were decreased from the operational values of (6e-4, 3e-4) to (2e-4, 2e-4) as the larger values were found to give a substantial warm bias in the global mid-troposphere. These lower values appear to be more consistent with the reduced spatial resolution used here. The model uses two stochastic physics schemes: Stochastically Perturbed Parametrization Tendencies (SPPT; Palmer *et al.*, 2009; Shutts *et al.*, 2011) and specific humidity perturbations (SHUM; Tompkins and Berner 2008), which perturb the humidity fields directly (Wang *et al.* 2019 give a description

TABLE A2 Summary of system details for 20CRv2c and 20CRv3.

	20CRv2c	20CRv3
Horizontal resolution	T62	T254
Effective resolution at equator	210 km	60 km
Vertical resolution	28 levels up to ~2.5 hPa	64 levels up to ~0.3 hPa
Temporal resolution	3–6 hourly	3 hourly
Forecast model	GFS2008EX	GFSv14.0.1
DA method	EnSRF with digital filtering	EnSRF with 4DIAU
Ensemble size	56	80
Observations	ISPDv3.2.9	ISPDv4.7
Sea ice fields	COBE-SST2	HadISST2.3
SST fields	SODAsi.2 (1851–2012): Reynolds OIv2 (2013–2014)	SODAsi.3 (20CRv3si. 1836-2012): HadISST2.2 (20CRv3mo, 1981–2015)
Ozone	Prognostic (McCormack *et al.*, 2006)	Specified (CMIP5; Cionni *et al.* 2011)

of the GFS implementation). The snow depth and lower three soil moisture levels are both subject to a 60-day relaxation to a monthly climatology (Saha *et al.*, 2010). The prescribed CO_2 and other physical parametrizations are unchanged from 20CRv2c and are described by Compo *et al.* (2011) and Saha *et al.* (2010) and summarized by Fujiwara *et al.* (2017).

The boundary conditions for 20CRv3, namely sea ice concentration and sea surface temperatures, have also been updated from those used in 20CRv2c. Sea ice concentrations for 20CRv3 are specified from HadISST2.3, which is identical to HadISST2.2 (Titchner and Rayner, 2014) from 1972 onwards. From 1850 to 1971, HadISST2.3 specifies sea ice extent from the Sea Ice Back To 1850 dataset (SIBT1850; Walsh *et al.* 2016). Prior to 1850, sea ice is specified as the 1860–1891 climatology of HadISST2.3. For SST fields, 20CRv3si prescribes eight members of pentad (interpolated to daily) SODAsi.3 SSTs duplicated ten times each for 1814 onward (Giese *et al.*, 2016). These fields are seasonally adjusted to the 1981–2010 HadISST2.2 climatology so that the 20CRv3si SST forcing climatology is the same as 20CRv3mo. HadISST2.1 climatology (1861–1890) is prescribed for the years 1813 and earlier. Regions where sea ice was ever indicated in HadISST2.3 are filled with HadISST2.2 daily (1963–2012), HadISST2.1 monthly interpolated to daily (1850-1962), or the 1861–1891 HadISST2.1 climatology (1849 and earlier). 20CRv3mo is distinguished by having prescribed SSTs from eight members of pentad (interpolated to daily) HadISST2.2 sea surface temperatures, duplicated ten times.

The operational GFS uses a hybrid 4DEnVar (Kleist and Ide, 2015) with a digital filter to prevent numerical instabilities from forming in the analysis step (Lynch and Huang, 1992; Huang and Lynch, 1993). The 20CRv2c system used the deterministic Ensemble Square Root Filter (EnSRF) (Whitaker and Hamill, 2002), as well as a digital filter. Assimilation cycles were completed every 6 hr; e.g. at 0000, 0600, 1200, and 1800 UTC. For example, the assimilation cycle completed at 0600 UTC includes an analysis valid at 0600 UTC and assimilates observations from 0300 to 0900 UTC. The digital filter would then be active for the first three forecast hours (i.e. from 0600 to 0900 UTC) and turned off for the next six forecast hours (i.e. 0900 to 1500 UTC). The cycle then repeats, assimilating observations from 0900 to 1500 UTC into the filtered 1200 UTC forecast. While the 20CRv3 system also utilizes the EnSRF, it includes a 4D Incremental Analysis Update (4DIAU; (Bloom *et al.*, 1996; Lei and Whitaker, 2016)) instead of the digital filter. Analysis cycles are still completed every 6 hr, assimilating observations in a 6-hr window centred at the analysis time (–3 hr and +3 hr from the analysis time.) In the context of the 4DIAU, assimilating observations results in three 'preliminary' analyses: one centred at the analysis time, one 3 hr prior, and one 3 hr after. The model is then re-run, starting 3 hr prior to the analysis time, with additional forcing added to guide the model towards the preliminary analyses but without the discontinuous jumps that arise from a typical ensemble Kalman filter step. The result of this run is the final analysis. The background, or forecast, is then generated by running the model for 6 hr without the IAU forcing from the end of the analysis window. In other words, to complete a cycle at 0600 UTC, preliminary analyses are calculated for 0300, 0600, and 0900 UTC; the model is re-run using the preliminary analyses as additional forcing to generate final analyses at 0600 and 0900 UTC, and finally, the model is run without IAU forcing from 0900 to 1500 UTC to generate the background fields for the next analysis cycle.

The pressure observations assimilated into 20CRv3 consist of an ACRE-facilitated blend of station data, marine observations, and tropical cyclone data merged to generate ISPDv4.7. The ISPDv3.2.9 used in 20CRv2c extended back to 1755 (Cram *et al.*, 2015), while ISPDv4.7 extends further back to 1722. The updated version includes over 4500 new station records from more than 30 collections, including

station data on six continents. Note that this includes new station records that are as short as one day, though many last for years or decades. Land station data are bias-corrected based on prior 60-day observation–first guess departures as in Compo *et al.* (2011); data that are missing an observation time are not included. The marine observations consist of ICOADS3+v2, including observations from OldWeather.org and the Australian Weather Detective project, as well as several additional ships recently digitized (Spencer *et al.* (2019); https://github.com/oldweather/ICOADS3.plus/releases – accessed 11 July 2019 – gives a description of processing procedures). ICOADS3+v2 fixes a date-misassignment present in ICOADS3 (Freeman *et al.*, 2017), and assigns local noon to reports without a defined hour of observation. Marine observations prior to 1870 are bias-corrected as described in Appendix B. The tropical cyclone data consist of IBTrACS V03r10 (Knapp *et al.*, 2010; Kruk *et al.*, 2010) combined with additional Pacific tropical cyclone data (Kubota, 2012). Recall that the observation errors assigned for each of these observations are given in Table 1. Note that the tropical cyclone data bypass the nonlinear quality control step (Appendix C).

Like 20CRv2c, 20CRv3 production was completed in parallel streams, starting from September of years 1834, 1839, 1844, 1849, ..., 2009. The first 16 months were taken as spin-up, and released data will start in January of 1836, 1841, 1846, ..., 2011. These streams were initialized from 20CRv2c ensembles.

B.1 | Pre-1870 ship bias correction

For 20CRv3, biases from ICOADS R3 and auxiliary marine observations taken prior to 1870 were precomputed and subtracted from the observations (Brohan, personal communication, 2017). Each ICOADS observation is associated with a 'deck' and an ICOADS ID; generally, the ICOADS ID is ship-specific, while the deck is a higher category consisting of collections of ships (Worley *et al.*, 2005; Woodruff *et al.*, 2011). For example, all Japanese whaling ships fall under the same deck, but different ships should have different ICOADS IDs. For each marine observation, an individual bias is calculated as the difference between the observed SLP and the climatological SLP for that time and location (calculated from a 1981–2010 climatology from 20CRv2c). Individual biases from each ICOADS identification for all years up to and including 1869 were aggregated and averaged. This average bias is subtracted from each observation with that ICOADS ID. At least ten observations were required to calculate a bias; any ICOADS IDs with fewer than ten observations total prior to 1870 were not bias corrected. Marine observations that did not have an ICOADS ID were aggregated by year and deck; thus for each year, all marine observations without an ICOADS ID that share a given deck ID received the same average bias correction.

C.1 | Nonlinear quality control algorithm

A nonlinear quality control (QC) algorithm that includes a representation of non-Gaussian observation errors has been implemented in the 20CRv3 system. Non-Gaussian observation errors are typically handled in variational DA systems with a procedure known as 'variational quality control' (VarQC; e.g. Anderson and Järvinen, 1999). Instead of assuming a Gaussian observation error, the distribution generally used is 'heavy-tailed' in order to represent gross instrument errors. The Kalman filter analogue of this approach was described first by Dharssi *et al.* (1992): the equations are identical to the familiar Kalman filter equations, except that the observation error variance depends on the observation innovation. The update is solved iteratively; within the serial EnKF (in which each observation is assimilated sequentially), each iteration is identical except that the observation error variance in each iteration is modified by dividing by the probability that the observation does not have a gross error, estimated using the innovation from the previous iteration (equations (15)–(17) in Dharssi *et al.*, 1992). The Huber norm observation error distribution is used (Tavolato and Isaksen, 2015), which implies that the probability of a gross error is zero for $|\mathbf{z}| <= c$ and $1 - |c/z|$ for $|z| > c$, where c is a threshold parameter and $\mathbf{z}$ is the innovation normalized by the observation error standard deviation: that is, $\mathbf{z} = (\mathbf{y} - \mathbf{Hx})/\sigma_{\mathrm{ob}}$, where $\mathbf{y}$ is the observation, $\mathbf{H}$ is the observation operator, $\mathbf{x}$ is the background model state, and σ_{ob} is the observation error standard deviation. In 20CRv3, c is set to a value of 1.1, similar to what is used for surface pressure observations in ECMWF operations. The Gaussian observation error standard deviation is reduced by an empirical factor $\min(1, 0.5 + 0.25c)$, following Tavolato and Isaksen (2015). In the 20CRv3 system, the observation-space update is iterated seven times to obtain the final observation-error variance estimate; this estimate is then used one final time to update the model state.

Effectively, the algorithm inflates the observation errors for observations that are inconsistent with nearby observations (within a localization radius), in a manner similar to the buddy-check algorithm used in the previous 20CRv2 system (Appendix B of Compo *et al.*, 2011). However, instead of a binary decision to accept or reject an observation, in the new procedure nearly all the observations are used, but those with a higher probability of having a gross error are assigned larger observation errors, and hence given less 'weight'.

Figure C1 shows the impact of the new QC algorithm for extratropical cyclone *Martin* in December 1999 (Ulbrich *et al.*, 2001). The model and set-up used are the same as in 20CRv2c, but with a higher resolution (T126). If only the

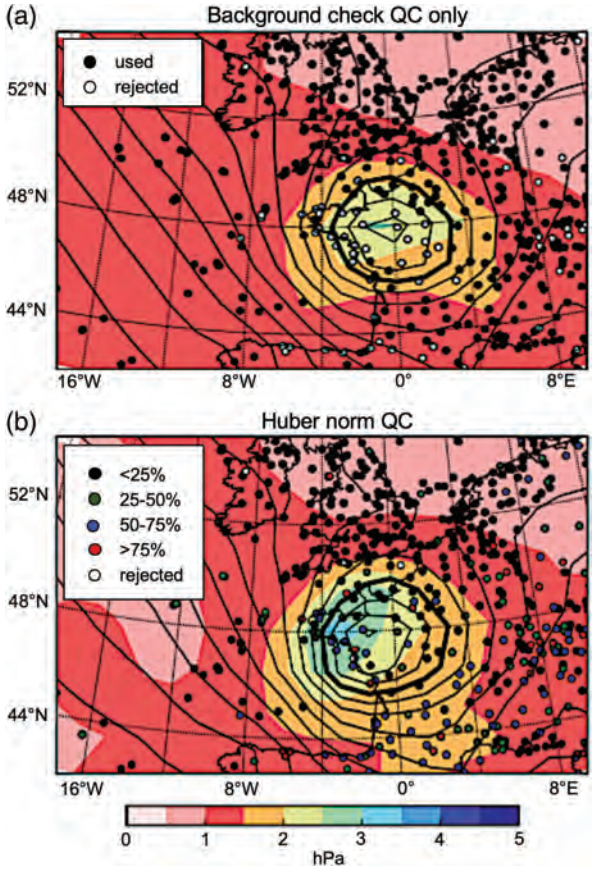

FIGURE C1 Analyzed ensemble mean surface sea-level pressure (solid contours; 980 hPa bold, 4 hPa interval) and background ensemble spread (shading, in hPa) for 1800 UTC on 27 December 1999, using (a) background check QC only and (b) the Huber norm QC. Circles represent observations; fill colours denote the probability of a gross error, if assimilated.

background check QC is used (Figure C1a), most of the observations within the cyclone and many over continental Europe are rejected. Recall from Section 4 that this QC rejects observations whose difference from the background field is larger than $3.2\sqrt{\sigma_b^2 + \sigma_{ob}^2}$; these observations are shown in white. If the background QC threshold is increased from 3.2 to 16 and the nonlinear QC is activated (Figure C1b), most of those observations are used, but given reduced weight. Figure C1b shows observations coloured by the probability of a gross error; a higher probability results in a smaller weight. As a consequence of the nonlinear QC, the ensemble mean cyclone is deeper and the background ensemble spread is larger in the vicinity of the cyclone. However, we found that the reduction of the background QC check threshold had an overall negative impact on analysis quality since suspect observations in data-sparse regions were not flagged by the nonlinear QC algorithm. We did find a small but significant positive impact of including the nonlinear QC algorithm in addition to the background QC check with a threshold of 3.2. This, in conjunction with whitelisting central pressure data from tropical cyclones, is the QC configuration used in the 20CR version 3 system.

D.1 | Adaptively varying localization length-scales in the serial EnKF

Unlike the fixed localization procedure in 20CRv2c, the version 3 system utilizes an adaptive localization scheme that avoids a specified value solely determining the localization length-scale. This scheme makes the localization length-scale proportional to the expected reduction of ensemble variance in observation space for each serially assimilated observation. Observations are sorted so the next observation assimilated is always the one with the largest expected ensemble variance reduction. This ensures that the localization length-scale decreases monotonically as observations are assimilated, so

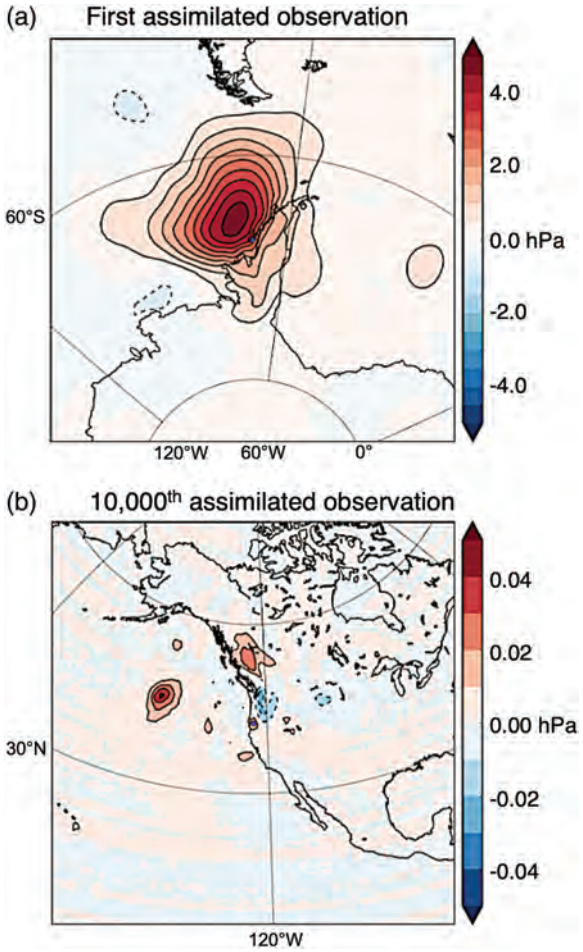

FIGURE D1 Surface pressure increment (hPa) for the (a) first (smallest value of ρ) and (b) 10,000th (ρ statistically indistinguishable from 1) observations assimilated globally on 1 April 2000. Note that the domains differ between (a) and (b) since the first assimilated observation is located west of the Antarctic peninsula, and the 10,000th assimilated observation is in the western US. Also note that the colour scales differ. A fixed localization length-scale of 4000 km in the horizontal and four scale heights in the vertical was used, and observations were assimilated in order of increasing ρ

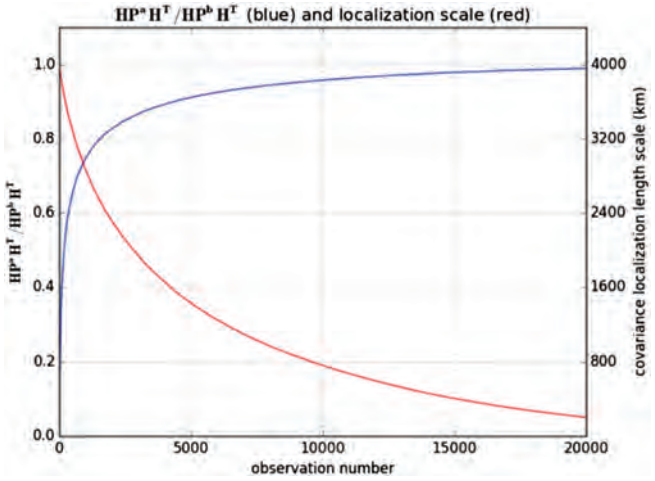

FIGURE D2 ρ (blue, left axis) and L (red, right axis), diagnosed using Equation D.2 with $r = 0.2$ and L_0=4000 km, as a function of observation number in the serial assimilation algorithm for the experiment described in Figure D1 but with adaptive localization

that the first (last) observation assimilated has the largest (smallest) expected variance reduction in observation space, and the largest (smallest) localization length-scale. Specifically, we hypothesize that for a single observation, the reduction of ensemble variance in observation space is inversely proportional to the optimal localization length-scale for that observation:

$$\rho = \mathbf{HP}^{a}\mathbf{H}^{T}/\mathbf{HP}^{b}\mathbf{H}^{T} = \mathbf{R}/(\mathbf{HP}^{b}\mathbf{H}^{T} + \mathbf{R}), \quad \text{(D1)}$$

where $\mathbf{H}$ is the linearized observation operator, $\mathbf{P}^{b}$ is the background ensemble covariance, $\mathbf{P}^{a}$ is the analysis ensemble covariance, and $\mathbf{R}$ is the original observation error variance. Small values of ρ correspond to a large reduction in variance of the observed variable when that observation is assimilated, which we assume implies a large signal-to-noise ratio in the ensemble covariance estimate, and therefore less severe localization (a longer localization length-scale) should be used. Conversely, as ρ asymptotes to unity, there is very little reduction in variance of the observed variable, implying that the ensemble covariance is noisy, requiring a more severe localization (a smaller localization length-scale) to be used. Empirically, we choose the following relationship between ρ and L, the localization length-scale:

$$L = L_0(1 - e^{-(1-\rho)/r}), \quad \text{(D2)}$$

where L is the localization length-scale, L_0 is the maximum allowed localization length-scale, and r is a parameter governing how tight the relationship between ρ and L is. The observations are assimilated serially in order of increasing ρ, and ρ is recomputed after each observation is assimilated.

To validate the underlying assumption that signal-to-noise ratio of the ensemble covariance estimate decreases with

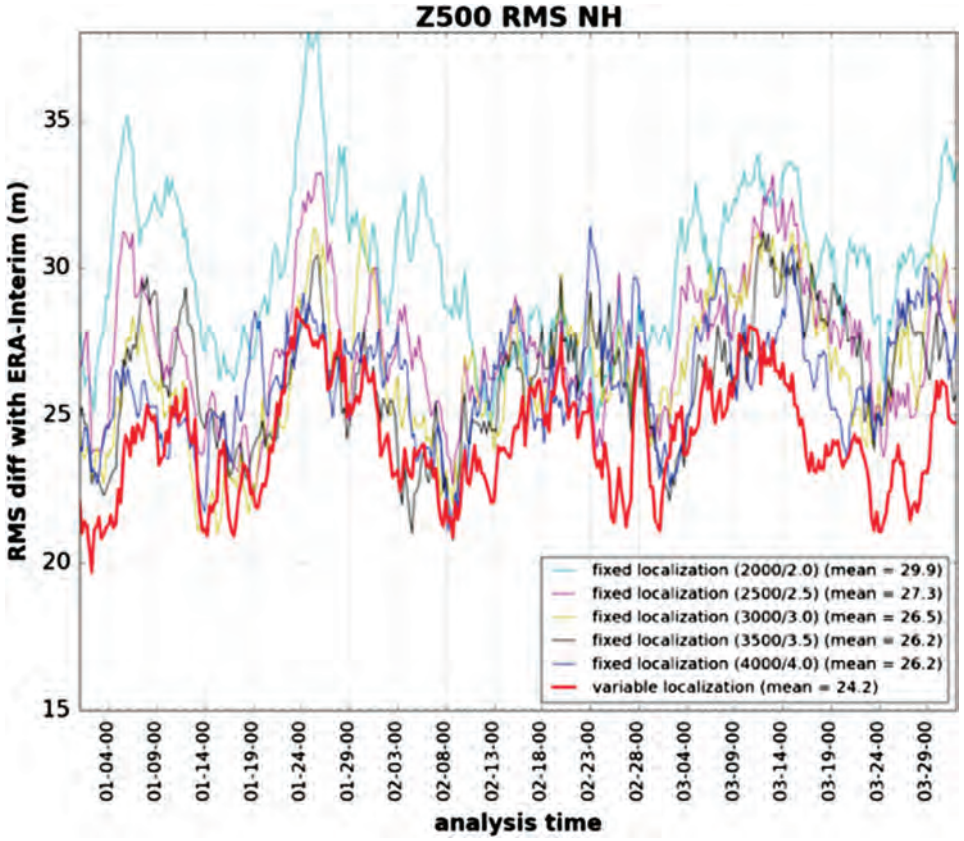

FIGURE D3 Time series of 500 hPa geopotential height error (relative to ERA-Interim reanalyses) for the experiments described in the text

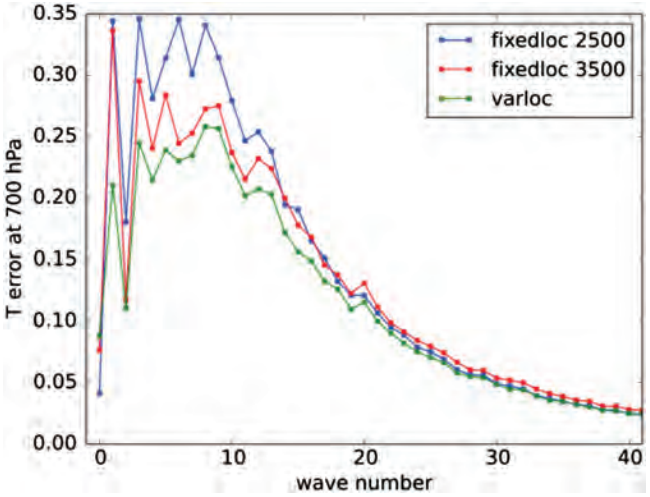

FIGURE D4 The error spectra of 700 hPa temperature for the adaptive localization experiment ('varloc') and two of the fixed localization ('fixedloc') experiments (2500/2.5 and 3500/3.5 km per scale height).

increasing ρ, we performed an experiment with a fixed localization length-scale L of 4000 km in the horizontal and four scale heights in the vertical in which the surface pressure observations were assimilated in order of increasing ρ. Figure D1 shows the surface pressure increment for the first and 10,000th observations assimilated in one cycle of this experiment.

As expected, the spatial coherence of the surface pressure increment generally decreases as a function of increasing ρ, and the increments become noisier and smaller scale as ρ increases. Figure D2 shows ρ as well as the localization length-scale L, diagnosed using Equation D.2 with $L_0 = 4000$ km and $r = 0.2$, as a function of the number of observations assimilated.

Based on these results, a series of experiments was performed for the period January–April 2000 to assess the impact of the adaptive localization scheme. Fixed localization experiments were run with horizontal and vertical localization scales corresponding to 2000/2, 2500/2.5, 3000/3, 3500/3.5 and 4000/4 (km and scale heights, respectively). A relaxation to prior spread coefficient of 0.9 was used throughout, station observations were bias corrected, and the QC algorithm described in Appendix C was used. These fixed localization experiments were compared to an adaptive localization experiment using the scheme described above with $r = 0.2$ and $L_0 = 4000$ km per four scale heights. Figure D3 shows time series of Northern Hemisphere geopotential height analysis errors at 500 hPa (relative to ERA-Interim) for all of the experiments.

The adaptive localization experiment shows a clear reduction in analysis error relative to all of the fixed localization experiments. The reason for the improvement can be understood using the error spectrum shown in Figure D4. For the fixed localization experiments, decreasing the localization length-scale reduces the analysis error at small scales, but increases the error at large scales. This is a well-known consequence of localization with a fixed length-scale in a system with multiple error scales (e.g. Buehner and Shlyaeva, 2015); in essence, tuning the single length-scale results in a compromise of balancing reductions in errors at some scales with increases in error at other scales. In contrast, the adaptive localization scheme is able to reduce errors at all scales simultaneously. Based on these tests, the 20CR version 3 system uses the adaptive localization scheme described here with $r = 0.2$ and $L_0 = 4000$ km per four scale heights.

This scheme has been tested with more complete observing systems than 20CR (including upper-air and satellite radiance observations), but there have been no significant positive impacts relative to using a fixed localization scale. We hypothesize that the homogeneity of the observing system used in 20CR (surface pressure observations only) is crucial. An observing system with many different types of observations (some of which may only be weakly correlated with model state variables) will require a more sophisticated metric for measuring sampling error than the expression for ρ given by Equation D.1.