ORIGINAL ARTICLE

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

[1]U. of CO Cooperative Institute for Research in Environmental Sciences, Boulder, CO, USA

[2]NOAA Earth System Research Laboratory, Boulder, CO, USA

Correspondence
Laura C. Slivinski, NOAA/ESRL Physical Sciences Division, 325 Broadway, Boulder,

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 in-

dividual 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 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 the effectiveness of 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 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
reanalysis, data assimilation, surface pressure

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; see Compo et al. (2006) and Compo et al. (2011) for 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 are available for assimilation. These reanalyses include the European Centre for Medium-Range Weather Forecasts (ECMWF) interim Re-Analysis 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 Re-Analysis 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 (see Fujiwara et al. (2017) for 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 tropical divergent circulation (Kinter III et al., 2004) and trends in temperature, integrated water vapor, 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, 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; see 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 will extend back to 1836 and will be released in 2019.

Historical reanalyses have broad areas of application because they span timescales 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 (Compo et al., 2013; Huang et al., 2016); climate dynamics (Huang et al., 2017); trends in hurricanes (Burn and Palmer, 2015), extra-tropical cyclones (Wang et al., 2013, 2016), and extremes in temperature and precipitation (Donat et al., 2016); blocking (Häkkinen et al., 2011; Rohrer et al., 2018); individual case studies of particular storms (Moore and Babij, 2017); historic climatology in remote regions (Lorrey and Chappell, 2016); El Niño (Giese et al., 2010; Deser et al., 2017); the Madden-Julian Oscillation (Klotzbach et al., 2016); convergence

zone activity (Lorrey et al., 2012; Harvey et al., 2019); seasonal and climatic responses to volcanic eruptions (Brohan et al., 2016; Paik and Min, 2017); weather typing (Jones et al., 2013, 2016); and the emerging field of decadal climate prediction (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 (Slivinski, 2018). Nevertheless, it is 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). 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 timescales, 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 improved the efficacy of the reanalysis. In other cases, version 3 benefited 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 were addressed, as well as other upgrades to the version 3 system. While results may strongly suggest some improvements were caused by particular changes to the system, many updates were made simultaneously. Thus, a single improvement in the 20CRv3 *dataset* can rarely be attributed to a specific change in the 20CRv3 *system*. Finally, this work is intended to review the 20CRv3 system itself; thus, results with the 20CRv3 dataset shown here will focus on a few 'test periods' between 1854 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 works.

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 to generate the datasets. 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. Results show that updates to the forecast model and data assimilation algorithm have improved the confidence estimation and reduced most of these biases in 20CRv3. In addition to addressing known issues, other developments in the version 3 system have resulted in further improvements. As shown in Section 4, updates to the localization procedure, quality control, and observation errors have improved the use of observations and resulted 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 background 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 the atmospheric state via the ensemble mean. Finally, surface pressure values are the only type of observations that are ever assimilated. While the temporal frequency, spatial density, and quality of these observations has 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. Feedback from reanalysis datasets that assimilate these observations can be used to improve this estimate in the future (e.g., see Figure 10 in Poli et al. (2013)). Below, specific details of the 20CRv2c and 20CRv3 systems are described.

20CRv2c consists of a 56-member ensemble of 3- to 6-hourly atmospheric fields at a spatial resolution of total spherical wavenumber 62 (about 2 deg.) with 28 vertical hybrid sigma-pressure levels. The model is based on an experimental version of the 2008 NCEP GFS, a spectral coupled atmosphere-land model with a comprehensive suite of physical parameterizations 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). It assimilated surface pressure from the International Surface Pressure Databank (ISPD) version 3.2.9 (Cram et al., 2015; Compo, 2015), which consists of station observations, marine observations from the International Comprehensive Ocean-Atmosphere Data Set (ICOADS, Worley et al. (2005); Woodruff et al. (2011)) version 2.5.P, 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 sea level pressure observations from the International Surface Database (ISD, Lott et al. (2008)). Procedures for blending the station component are described by Yin et al. (2008). The observations were assimilated via a deterministic, square-root ensemble Kalman filter (EnKF) (Whitaker et al., 2004; Compo et al., 2006; Whitaker and Hamill, 2002). The sea surface temperatures (SSTs) were prescribed using the fields from the Simple Ocean Data Assimilation system with sparse observational input version 2 (SODAsi.2; Giese et al. (2016)), which consists of 18 pentad realizations that used 20CRv2 atmospheric fields as boundary forcing, tapered to COBE v2 SSTs poleward of 60 degrees N and S (Giese et al., 2016; Hirahara et al., 2014). A misspecification in the sea ice concentration for 20CRv2 was fixed in 20CRv2c, which used COBE v2 sea ice concentration down to fractions of 0.15 (Hirahara et al., 2014). All other components of the system are only constrained by the assimilation of pressure observations. More details on the 20CRv2c system are given in Appendix A.

To address significant issues in the 20CRv2c dataset, and as a result of general progress in the fields of modeling and data assimilation, several aspects of the 20CR system were updated before producing 20CRv3. Broadly, 20CRv3 benefits from an improved, higher-resolution model; a larger observational database; updated data assimilation methods; and a larger ensemble size. The atmospheric model used in 20CRv3 was updated to the 2017 version of the NCEP GFS with a resolution of total spherical wavenumber 254 (about 0.5 deg. horizontal resolution) and 64 vertical hybrid sigma-pressure levels; differences between the version of the GFS operational in fall 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. In version 3, the dry air pressure was constrained; see Section 3.

The 20CRv2c dataset began in 1851 since the COBE sea ice fields were first available in 1850; this year was considered 'spin-up'. 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 with appropriate boundary conditions. 1804 is the first year that every 6-hour 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.

20CRv3 will 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. Of the 10 members available, two of the ensemble members had quite different bias adjustments from the others; thus, 20CRv3mo only uses the remaining 8 members as boundary conditions. 20CRv3si prescribes SSTs from the pentad 8-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, the 8-member SST ensemble was duplicated 10 times to create 80 members. 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. (2015, updated 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. The new observational dataset, ISPD version 4.7, blends ISD version 3.0 with additional station observations, archived and previously undigitized terrestrial data submitted to the ISPD from international ACRE partners, IBTrACS version 3, and 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, Old-Weather.org, and the Australian Weather Detective project (see https://github.com/oldweather/ICOADS3.plus/releases 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 4-dimensional incremental analysis update (Bloom et al., 1996; Lei and Whitaker, 2016) and no digital filtering; see Section 3 and Appendix A. Additionally, 20CRv2c interpolated station pressure observations to the model surface prior to assimilation, while 20CRv3 assimilates them at the observation level and absorbs the vertical interpolation of the background forecast into the observation operator (**H**). As will be discussed in Section 4, 20CRv3 includes a nonlinear quality control 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 (see Appendices B-D for 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 (see Compo et al. (2011), their Appendix B, for more details). Finally, the baseline observation errors used in 20CRv3 are given in Table 1. Column 4 ('station') refers to observations of surface pressure, while column 5 ('SLP only') refers to stations that only reported pressure reduced to sea level. Observation errors are increased by 0.001 hPa per 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 (see Table A.1).

3 | ADDRESSING ISSUES IN 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 Feb 2019). While useful, the 20CRv2 dataset has several issues including a misspecification of polar sea ice that resulted in warm 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 (described briefly in Giese et al. (2016) and in detail here) 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, limitations of it became apparent. Simultaneously, the many studies using 20CRv2c motivated further data rescue efforts, and the amount 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-hour window of 20CRv2c (solid black) and 20CRv3 (dashed gray). 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 quality control system (see Section 4 and Appendix C for details.)

A new version of the 20CR system could make use of this growing set of observations, as well as general progress in modeling and data assimilation 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 sea level pressure, 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 dataset. A defining characteristic of 20CR is its use of an ensemble data assimilation 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. As an example, Figure 2 includes time series of uncertainty in sea level pressure over the zonal band from 65°S to 40°S. This region is particularly important for investigations of anthropogenic CO_2 uptake, and long time series are needed for studies of its decadal variability (eg, Landschützer et al. (2015)). The analyzed sea level pressure ensemble spread is plotted (thick red curve) along with the temporal spread of the analyzed sea level pressure ensemble mean (thin orange curve) and the number of observations assimilated per 6-hour window (black curve) in this region. This region has relatively few observations available (compare with Fig. 1), and the effects of World War I (1914–1918) and World War II (1939–1945) are particularly striking. The temporal spread is calculated as the standard deviation of the ensemble mean across a centered 61-day window, and all time series have a 1-year running average applied. 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.

Ensemble spread is only an estimate of uncertainty, though, and it is not always reliable. For instance, a well-known issue with the EnKF is the tendency for the 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 solution to 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 (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 the foundation of the International Meteorological Committee in 1873, and 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 (Fig. 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.

Another issue demonstrated by Fig. 2 is the under-confidence of the ensemble spread. The temporal spread (thin orange curve) 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} determined jointly by a relaxation parameter p_{relax} and the density of the observations in that region. p_{relax} can vary from 0 (no inflation) to 1 (inflate to prior spread). Initial tests with the 20CRv3 system used 0.9 globally. These tests (not shown) suggested that $p_{relax} = 0.9$ was too large in the Southern Hemisphere. Thus, the final 20CRv3 system has values of $p_{relax} = 0.9$ for 20°S–90°N and $p_{relax} = 0.7$ for 90°S–30°S. In the transition zone 30°S–20°S, p_{relax} varies linearly from 0.7 to 0.9. λ_{inf} varies in both space and time, and is calculated at each assimilation step. 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 increases the ensemble spread. This effect is particularly strong in the tropics, which reduces the need for inflation in this region (e.g., Fig. 3d). Outside of 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 Fig. 3 and Table 2).

Figure 4 illustrates the boreal winter confidence in the early 20th century from versions 2c and 3 of 20CR. Here,

'confidence' is defined as the difference of the normalized time-averaged ensemble standard deviation from 1: $conf = 1 - spread_{ens}/spread_{clim}$, where $spread_{ens}$ is the time-averaged standard deviation of the ensemble of analyzed SLP from the stated version of 20CR, and $spread_{clim}$ is the temporal standard deviation of the 20CRv2c ensemble mean 6-hourly SLP over Jan-Feb-Mar from 1981-2010. In other words, $spread_{clim}$ 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 Fig. 4a-b) denotes 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 Fig. 4c in this region), which was previously more uncertain than a climatological mean 'analysis'. There is also more spread over the densely-observed regions of North America and Europe (blue shading in Fig. 4c in these regions). Results (not shown) using independent observations from U.K. 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. 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, a sea level pressure 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 sea level pressure (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 2008 GFS model 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 Fig. 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; Ansell et al., 2006; Allan and Ansell, 2006), combined with the 20CRv2c system allowing the global dry pressure to be updated during the assimilation cycle. 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. Panel (b) differs from (a) in that it used an experiment that assimilated about 10% fewer ship observations than 20CRv2c, but with an otherwise identical setup. Overall, the anomalies are less negative. Thus, assimilating more ship observations from 1851–1853 negatively biased the globally-averaged analyzed sea level pressure by as much as 4 hPa.

Two improvements in 20CRv3 have addressed this issue. First, the global dry 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 sea level pressure 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 (see Appendix B for more details).

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

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 (Fig. 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 (see, for example, van Loon and Jenne (1972); Raphael (2004)) and would be erroneously removed by the bias correction scheme (Fig. 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., Fig. 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. Once example stems from the use of a digital filter in the forecast step of 20CRv2c (Lynch and Huang, 1992; Huang and Lynch, 1993) 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 hours 3-6. 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 the time steps 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 total model fields, thereby eliminating the spurious tendency trends seen during the forecasts of version 2c. Figure 8b 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 hour and 0-3 hour forecasted precipitation rates) from 20CRv2c fields (Fig. 8a) is 0.72 mm/day, while the average from 20CRv3 (Fig. 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 year 2002 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 10m wind fields (Figure 10). While it is possible that this is due to an error in the process of converting between the spectral model space and the physical grid space, a more likely cause is an error in the surface height field used in 20CRv2c. In particular, the spectral transform of a higher-resolution height field to the lower-resolution field used as an input to the model may have been incorrect, leading to a similar spectral signature in the 20CRv2c height field (not shown). However, the version 3 system does not suffer from this issue: 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, 20CRv3 benefited from other updates to the system that led to overall improvements. In particular, the version 3 system uses adaptive quality control 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, have resulted in smaller forecast errors, large-scale reductions in model bias, and more accurate representations of storms.

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 data assimilation 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 cutoff 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 Gaspar-Cohn function goes to zero. The optimal value for the localization length scale may be a function of many aspects of the data assimilation 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 4000 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 to localization, a five-step quality control (QC) process was employed in 20CRv2c. Briefly, this consisted of: (1) a gross error check that removed sea level pressure observations outside the plausible range of 880 to 1060 hPa; (2) a background check that flagged observations that were too far from the first guess value; (3) a buddy check to determine whether assimilating each observation improves or degrades the fit of the analysis to nearby observations; (4) a thinning step, which ignored observations that would not significantly decrease the spread of the analysis if assimilated; and finally, (5) a bias correction to the station observations. More details on this quality control process are found in Appendix B of Compo et al. (2011).

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 quality control procedure jointly with an adaptive localization algorithm. Observations must still pass steps (1) and (2) above: if the observation is outside a plausible range or if the observation is too far from the first guess, the observation is rejected. In 20CRv3, the first step will reject observations that are outside the range 850 and 1090 hPa (note this is less strict than the first step used in 20CRv2c). The second step will reject an observation if the absolute difference between it and the first guess is greater than $3.2 * \sqrt{\sigma_b^2 + \sigma_{ob}^2}$, 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 {steps (3) and (4) above} to reject observations that do not decrease the analysis ensemble spread. Instead, the adaptive quality control assigns these observations larger errors and smaller localization radii, so that their region of influence is essentially zero. Details of the adaptive quality control and localization procedures used in 20CRv3 can be found in Appendices C-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 4000 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 longer localization length scales. Observations located within areas that are sparsely observed (such as the Southern Ocean and Antarctica) still have fairly long 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 that have 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 be able to take advantage of these observations to the greatest extent possible, particularly when the observation network is fairly sparse. Here we show that, while 20CRv2c performs fairly well under many measures, the updated algorithms used in 20CRv3 produce clear improvements.

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. These expected errors are defined as the root-mean of the sum of the observation error variance σ_{ob}^2 and the spread of the first guess ensemble interpolated to the observation time and location σ_b^2. The 'actual' errors are the root-mean-square of the difference

between the observed pressure and the first-guess pressure field interpolated to the 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. We consider time series of regionally- and annually-averaged forecast errors from 20CRv2c for the Northern Hemisphere (20°N to 90°N), Tropics (20°S to 20°N), and Southern Hemisphere (90°S to 20°S). These errors are averaged over all observations in the given region for each year and plotted in Figure 12, along with the number of assimilated observations per year (blue; right-hand axis).

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 match the expected errors fairly well in all decades after 1860, but the error in the tropics seems to show effects of changes in observing network more strongly than in the Southern or Northern Hemispheres. With a few exceptions, the tropical errors are smaller than expected by the 20CRv2c system. Conversely, the 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 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.

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 (Fig. 7), the ship bias correction provides further improvement. The overall result of the new algorithms in the version 3 system is that RMS errors have decreased overall in several different test periods. Figure 13 shows the actual and expected errors in 1854 (when the marine observation bias correction was applied), 1935, and 2000. The actual errors in 20CRv3 are lower than the actual errors in 20CRv2c in almost all cases, and the difference between the actual and expected errors is often smaller in v3 than in v2c. While the actual errors in 1854 are still larger than the expected errors in 20CRv3, this difference is smaller than in 20CRv2c. These years are generally representative of their respective time period; however, there is a period of time surrounding World War I when the actual error from 20CRv3 is still larger than the expected error (not shown).

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 from each station and the first-guess pressure at that location. Ideally, the consistent mismatch between observations and the model first guess are biases in the observations that are removed; however, the 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). Conversely, small-scale biases may point to a mis-assignment of station elevation or position: these are observational biases that should be corrected.

Figure 14a-c 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 (see 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 (Fig. 14d-f). This may suggest that the model used in version 3 is less biased than in 20CRv2c. 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; see the Himalaya) and nationality (due

to country-specific calibration and correction methods; c.f. 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 15) and 500 hPa geopotential height fields (Figure 16) that are closer to those of ERA-Interim and JRA-55, particularly in northern Eurasia.

4.3 | Representation of major storms

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 associate 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; see Compo et al. (2011); Knapp et al. (2010) for more details. Since these data are often significantly lower than the nearby station observations, they would generally fail the 'buddy check' step of quality control that compares observations to their nearest neighbors (see Section 4.1 above and Appendix B of Compo et al. (2011)); therefore, the 20CR system has IBTrACS data bypass these checks and assimilates these deep-low data.

In version 2c, these data were assigned large observation errors (much higher than any other type of observation) to prevent numerical instabilities from arising immediately after assimilation; see Table A.1. 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, however, can use these observations more effectively. Assimilating TC low-pressure values into the version 3 system, with an updated, higher-resolution forecast model and 4DIAU, does not generate instabilities, and so the IBTrACS data can be assigned smaller observation errors (see 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 17 illustrates the analyzed sea level pressure fields (contours) from 4 reanalyses, as well as locations of observations available between 16 Aug 1915 2100 UTC and 17 Aug 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. Observations that were assimilated are shown as solid circles, while observations that were rejected by the QC system are open circles. Blue circles represent station and marine observations, and red circles represent IBTrACS data. 20CRv2c (panel a) assimilated the IBTrACS report of 940 hPa on 17 Aug 1915 at 0600 UTC, producing an analyzed value of 986 hPa at the center of the storm. In version 3 (panel b), the storm is even stronger, and the analyzed value at the center of the storm has decreased to 962 hPa, reducing the observation-analysis departure calculated from 20CRv2c by half. For comparison, ERA-20C (panel c) rejected the IBTrACS observations during quality control and analyzed a low pressure system that is weaker than that of 20CRv2c and has a misaligned center. The CERA-20C system also rejected the IBTrACS observations, but assigned larger errors to the nearby station data than ERA-20C (Laloyaux et al., 2018), thereby showing no trace of the storm

(panel d). Laloyaux et al. (2018) conducted experiments with the CERA-20C system in which these observations were white-listed, and found this yielded better performance than CERA-20C for two hurricanes in 1900 and 2005 (their Fig. 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.

Figure 7 and Figure 15a-b demonstrate another potential issue with the Southern Hemisphere: a high pressure bias over Antarctica. Figure 7 shows that the 1854 annual average sea level pressure 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 15 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 sea level pressure 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 (Fig. 15c-d) is the opposite sign in this region.

A third Southern Hemisphere issue, regarding a trend in sub-Antarctic sea level pressure, was first brought to light during an investigation of ERA-20C (Poli et al., 2015, 2016). This dataset has a high pressure bias south of 60°S in the early 20th century that is particularly strong in austral summer (Dec-Jan-Feb). Comparisons with 20CRv2c show that it has a similar bias as ERA-20C, though it is not as strong in DJF. Figure 18 shows the seasonal time series of sea level pressure area-averaged poleward of 60°S for several datasets including 20CRv2c, ERA-20C, CERA-20C, and ERA-Interim. Data from an ensemble of model simulations using the 20CRv2c system but that did not assimilate any observations ('no DA') is also included, as well as preliminary 20CRv3 data for the test periods 1910-1930 and 1990-2010. There is a drop and subsequent increase in SLP from 20CRv2c in all seasons (most notably in Sept-Oct-Nov) from 1890-1910, with another significant drop-off between 1940 and 1960. The preliminary 20CRv3 data agrees 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 spread to the unobserved, and thus unconstrained, region farther south. They argue that this is due to assigning observation errors that are too small; in CERA-20C, the errors varied in time, and larger values were assigned in this time period. This has significantly decreased the trend in SLP from CERA-20C in all seasons, though it remains somewhat in DJF, when the bias in ERA-20C was most obvious. Hegerl et al. (2018) point out that the years 1906-1915 also exhibit anomalously cold SSTs in the Southern Ocean (their Fig. 2c), possibly due to instrumental biases; investigations into this issue and possible connections with the SLP signals in Fig. 18 are ongoing.

6 | CONCLUSIONS

With the growing need to understand and predict climate and extreme weather variations on decadal to centennial timescales, 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. It is, therefore, becoming more important that these reanalyses are 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 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 data assimilation algorithms, observation processing, and quality control; and an updated ISPD observation database. Several other issues with 20CRv2c exist that have not been discussed here, including spinup 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. In addition, there are high pressure biases over Antarctica and the Southern Ocean throughout time that may or may not be real signals; gathering high-quality independent observations to use for validation 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 challenge. 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, 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/group/fv3gfs); the changes resulting from this model versus the previous spectral model need to be investigated. Recent investigations into coupled data assimilation algorithms, and the first implementation of a quasi-strongly coupled data assimilation 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 data assimilation 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 (Univ. of New South Wales), M. Barriendos (Univ. of Barcelona), T. Brandsma (KNMI), Y. Brugnara (Univ. of Bern), O. Bulygina (All-Russia Research Institute of Hydrometeorological Information), A. Dawson (Univ. of Aberdeen), J. Filipiak (Univ. of Gdansk), P. Groisman (NC State University Research Scholar at NOAA NCEI), J. Holopainen (Univ. 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), O. 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). 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 Support 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) on an earlier version 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 (Grants No. 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, quality control and analyse early instrumental meteorological data across the world. The following people are financially supported

by the University of Giessen and digitized 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, K. Nan, C. Samaras, A. Tsikerdekis, C. Athanasiou. L. Dergianli is acknowledged for quality control, coordination, management and preparation of all Uni. Giessen data. S. Jourdain, M. A. Valente, M. Brunet, J. Luterbacher, R. 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. Allan acknowledge Climate Science for Service Partnership China Project (CSSP).

GPCP data provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their web site at https://www.esrl.noaa.gov/psd/. The ERA-Interim, ERA-20C, and CERA-20C datasets are courtesy of ECMWF. 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.

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). *J. Hydrometeor.*, **4**, 1147–1167.

Allan, R. 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. URL: https://doi.org/10.1175/JCLI3937.1.

Allan, R., 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. URL: 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: Nonlinear Phenomena*, **230**, 99 – 111. URL: http://www.sciencedirect.com/science/article/pii/S0167278906002168. Data Assimilation.

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.

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. URL: 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: Atmospheres*, **109**. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2004JD004536.

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**.

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. URL: 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. URL: 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 Past Discuss*, **2016**, 1–11.

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. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/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. URL: 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**, 12838.

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., 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. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/grl.50425.

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. et al. (2011) The twentieth century reanalysis project. *Quarterly Journal of the Royal Meteorological Society*, **137**, 1–28.

Compo, G. P. e. a. (2015) The International Surface Pressure Databank version 3. http://dx.doi.org/10.5065/D6D50K29. Accessed 12 Sept 2018.

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.

Daley, R. (1993) *Atmospheric data analysis*. No. 2. Cambridge University Press.

Dee, D. P., Uppala, S., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, d. P. et al. (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. URL: http:https://doi.org/10.1256/qj.05.108.

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., Manabe, T., Parker, D., Reynolds, R. and Woodruff, S. (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**, 11,174–11,189. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016JD025480.

Ferguson, C. R. and Villarini, G. (2012) Detecting inhomogeneities in the Twentieth Century Reanalysis over the central United States. *Journal of Geophysical Research: Atmospheres*, **117**. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2011JD016988.

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., 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. URL: https://www.atmos-chem-phys.net/17/1417/2017/.

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**, 177–183,148. URL: https://search.proquest.com/docview/232619945?accountid=28266.

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: Oceans*, **121**, 6891–6910. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016JC012079.

Häkkinen, S., Rhines, P. B. and Worthen, D. L. (2011) Atmospheric blocking and Atlantic multidecadal ocean variability. *Science*, **334**, 655–659. URL: http://science.sciencemag.org/content/334/6056/655.

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.

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 20th Century Reanalysis. *Monthly Weather Review*, **0**, null. In press.

Hegerl, G. C., Brönnimann, S., Schurer, A. and Cowan, T. (2018) The early 20th century warming: Anomalies, causes, and consequences. *Wiley Interdisciplinary Reviews: Climate Change*, **9**, e522. URL: https://onlinelibrary.wiley.com/doi/abs/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.

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

— (2001) A sequential ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review*, **129**, 123–137.

747 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*,
748 **46**, 1131–1150. URL: https://doi.org/10.1007/s00382-015-2636-8.

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

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

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

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

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

759 Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J. et al. (1996)
760 The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, **77**, 437–472.

761 Kent, E. C., Fangohr, S. and Berry, D. I. (2013) A comparative assessment of monthly mean wind speed products over the global
762 ocean. *International Journal of Climatology*, **33**, 2520–2541. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.
763 1002/joc.3606.

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

766 Kistler, R., Kalnay, E., Collins, W., Saha, S., White, G., Woollen, J., Chelliah, M., Ebisuzaki, W., Kanamitsu, M., Kousky, V. et al.
767 (2001) The NCEP–NCAR 50-year reanalysis: monthly means CD-ROM and documentation. *Bulletin of the American Mete-*
768 *orological society*, **82**, 247–268.

769 Klotzbach, P. J., Oliver, E. C. J., Leeper, R. D. and Schreck, C. J. (2016) The relationship between the Madden-Julian Oscillation
770 (MJO) and southeastern New England snowfall. *Monthly Weather Review*, **144**, 1355–1362. URL: https://doi.org/10.
771 1175/MWR-D-15-0434.1.

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

774 Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., Onogi, K., Kamahori, H., Kobayashi, C., Endo, H. et al. (2015)
775 The JRA-55 reanalysis: General specifications and basic characteristics. *Journal of the Meteorological Society of Japan. Ser. II*,
776 **93**, 5–48.

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

779 Laloyaux, P., de Boisseson, E., Balmaseda, M., Bidlot, J.-R., Broennimann, S., Buizza, R., Dalhgren, P., Dee, D., Haimberger, L.,
780 Hersbach, H. et al. (2018) CERA-20C: A coupled reanalysis of the Twentieth Century. *Journal of Advances in Modeling Earth*
781 *Systems*, **10**, 1172–1195.

782 Landschützer, P., Gruber, N., Haumann, F. A., Rödenbeck, C., Bakker, D. C. E., van Heuven, S., Hoppema, M., Metzl, N., Sweeney,
783 C., Takahashi, T., Tilbrook, B. and Wanninkhof, R. (2015) The reinvigoration of the Southern Ocean carbon sink. *Science*,
784 **349**, 1221–1224. URL: http://science.sciencemag.org/content/349/6253/1221.

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

van Loon, H. and Jenne, R. L. (1972) The zonal harmonic standing waves in the southern hemisphere. *Journal of Geophysical Research*, **77**, 992–1003. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JC077i006p00992.

Lorenc, A. C. (1986) Analysis methods for numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, **112**, 1177–1194. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/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, N., Vose, R., Del Greco, S., Ross, T., Worley, S. and Comeaux, J. (2008) The integrated surface database: partnerships and progress. In *Extended Abstracts, 24th Conf. on Interactive Information and Processing Systems*.

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

Middleton, K. (1964) *The history of the barometer*. Johns Hopkins University Press.

Monmonier, M. (1999) *Air apparent: How meteorologists learned to map, predict, and dramatize weather*. University of Chicago Press.

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.

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. URL: 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.

Poli, P., Hersbach, H., Berrisford, P. and other authors (2015) ERA-20C Deterministic. Shinfield Park, Reading.

Poli, P., Hersbach, H., Dee, D. P., Berrisford, P., Simmons, A. J., Vitart, F., Laloyaux, P., Tan, D. G., Peubey, C., Thépaut, J.-N. et al. (2016) ERA-20C: An atmospheric reanalysis of the twentieth century. *Journal of Climate*, **29**, 4083–4097.

Poli, P., Hersbach, H., Tan, D. G. H., Dee, D., Thépaut, J.-N., Simmons, A., Peubey, C., Laloyaux, P., Komori, T., Berrisford, P., Dragani, R., Trémolet, Y., Hólm, E. V., Bonavita, M., Isaksen, L. and Fisher, M. (2013) The data assimilation system and initial performance evaluation of the ECMWF pilot reanalysis of the 20th-century assimilating surface observations only (ERA-20C). 59. URL: https://www.ecmwf.int/node/11699.

Raphael, M. N. (2004) A zonal wave 3 index for the Southern Hemisphere. *Geophysical Research Letters*, **31**. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/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. *Journal of Climate*, **19**, 446–469. URL: 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. URL: https://doi.org/10.1175/JCLI-D-17-0350.1.

825 Slivinski, L. C. (2018) Historical reanalysis: What, how, and why? *Journal of Advances in Modeling Earth Systems*, **10**, 1736–1739.
826 URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018MS001434.

827 Slonosky, V. C. and Graham, E. (2005) Canadian pressure observations and circulation variability: Links to air temperature.
828 *International Journal of Climatology: A Journal of the Royal Meteorological Society*, **25**, 1473–1492.

829 Titchner, H. A. and Rayner, N. A. (2014) The Met Office Hadley Centre sea ice and sea surface temperature data set, ver-
830 sion 2: 1. Sea ice concentrations. *Journal of Geophysical Research: Atmospheres*, **119**, 2864–2889. URL: https://agupubs.
831 onlinelibrary.wiley.com/doi/abs/10.1002/2013JD020316.

832 Trenberth, K. E. and Smith, L. (2005) The mass of the atmosphere: A constraint on global analyses. *Journal of Climate*, **18**,
833 864–875.

834 Walsh, J. E., Chapman, W. L. and Fetterer, F. (2015, updated 2016) Gridded monthly sea ice extent and concentration, 1850
835 onward, version 1. Boulder, Colorado USA. NSIDC: National Snow and Ice Data Center.

836 Wang, X. L., Feng, Y., Chan, R. and Isaac, V. (2016) Inter-comparison of extra-tropical cyclone activity in nine reanalysis datasets.
837 *Atmospheric Research*, **181**, 133 – 153.

838 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
839 variability of extra-tropical cyclone activity in the ensemble of twentieth century reanalysis. *Climate Dynamics*, **40**, 2775–
840 2800.

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

843 Whitaker, J. S. and Hamill, T. M. (2002) Ensemble data assimilation without perturbed observations. *Monthly Weather Review*,
844 **130**, 1913–1924.

845 — (2012) Evaluating methods to account for system errors in ensemble data assimilation. *Monthly Weather Review*, **140**, 3078–
846 3089.

847 Woodruff, S. D., Worley, S. J., Lubker, S. J., Ji, Z., Eric Freeman, J., Berry, D. I., Brohan, P., Kent, E. C., Reynolds, R. W., Smith, S. R.
848 et al. (2011) ICOADS Release 2.5: extensions and enhancements to the surface marine meteorological archive. *Interna-*
849 *tional journal of climatology*, **31**, 951–967.

850 Worley, S. J., Woodruff, S. D., Reynolds, R. W., Lubker, S. J. and Lott, N. (2005) ICOADS release 2.1 data and products. *Interna-*
851 *tional Journal of Climatology*, **25**, 823–842. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.1166.

852 Yin, X., Gleason, B., Compo, G., Matsui, N. and Vose, R. (2008) The International Surface Pressure Databank (ISPD) land compo-
853 nent version 2.2. *National Climatic Data Center: Asheville, NC. Available from ftp://ftp. ncdc. noaa. gov/pub/data/ispd/doc/ISPD2*,
854 **2**, 2006–2100.

855 Zhang, L., Kumar, A. and Wang, W. (2012) Influence of changes in observations on precipitation: A case study for the Cli-
856 mate Forecast System Reanalysis (CFSR). *Journal of Geophysical Research: Atmospheres*, **117**. URL: https://agupubs.
857 onlinelibrary.wiley.com/doi/abs/10.1029/2011JD017347.

TABLE 1 Platform-dependent baseline observation errors in the 20CRv3 systems (in hPa). 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.

TABLE 2 Covariance inflation parameters used in 20CRv2c as a function of latitude and year. A value of 1 corresponds to no inflation. NH=90°N–30°N; Tropics = 30°N–30°S; SH = 30°S–90°S.

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, Northern Hemisphere; SH, Southern Hemisphere.

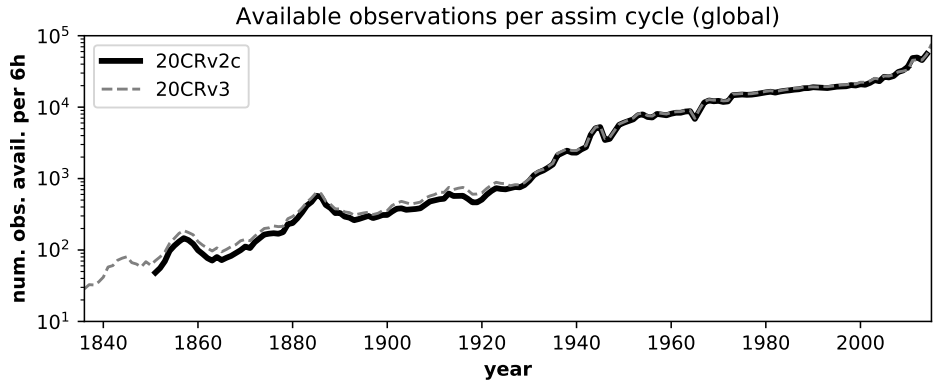

FIGURE 1 Time series of the annual average number of observations available to be assimilated globally per 6-hour window within 20CRv2c (ISPDv3.2.9, solid black) and 20CRv3 (ISPDv4.7, dashed gray).

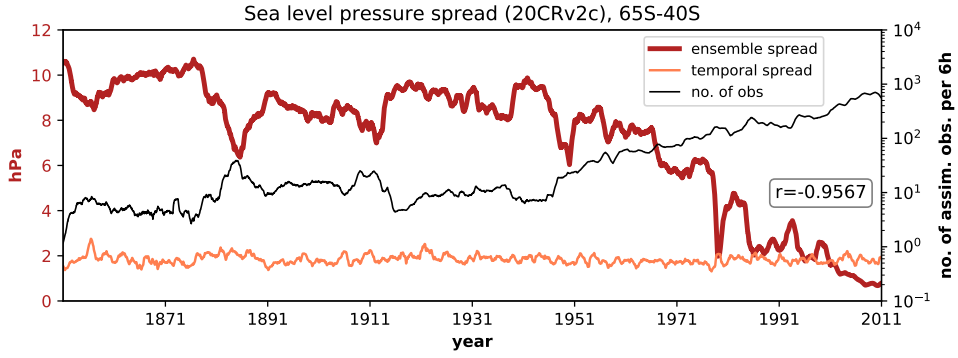

FIGURE 2 Time series of ensemble spread (thick dark red curve) and temporal spread of the ensemble mean (thin orange curve) for sea level pressure from 20CRv2c averaged over the zonal band from 65°S to 40°S. Number of observations assimilated per 6-hour window in this region is shown in black. A 1-year 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.

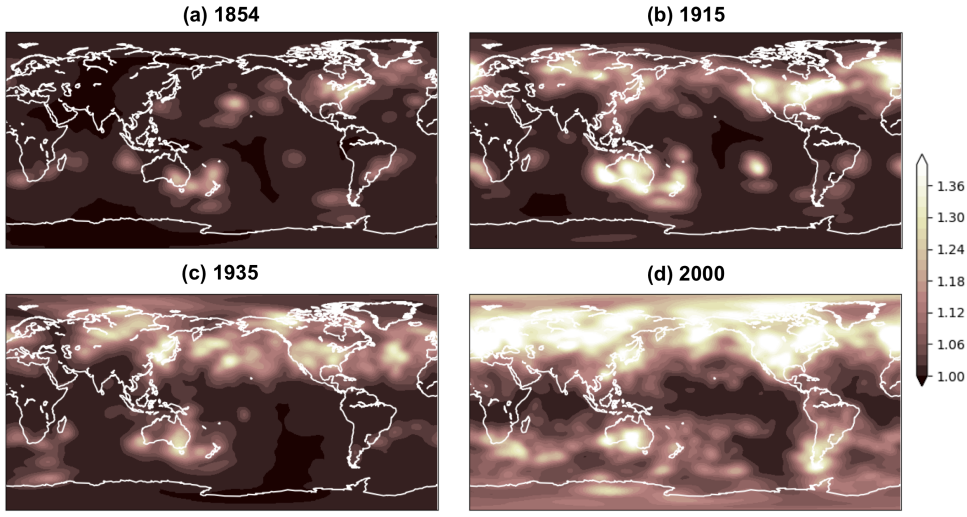

FIGURE 3 Adaptive inflation parameter λ_{inf} (unitless) used in 20CRv3 system for 0000 UTC on (a) 1 Sept 1854; (b) 1 June 1915; (c) 10 Feb 1935; (d) 1 Sept 2000. A value of 1 corresponds to no inflation.

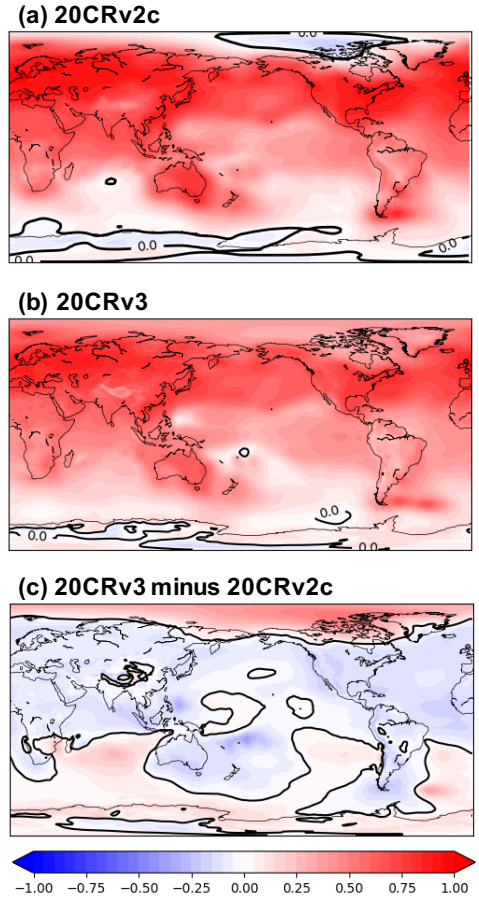

FIGURE 4 Normalized confidence of SLP averaged over JFM for 1916-1918 from (a) 20CRv2c and (b) 20CRv3, as well as (c) the difference (20CRv3 minus 20CRv2c). In (a)-(b), zero (black contour) represents climatological uncertainty, blue represents less certainty than climatology, and red represents more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

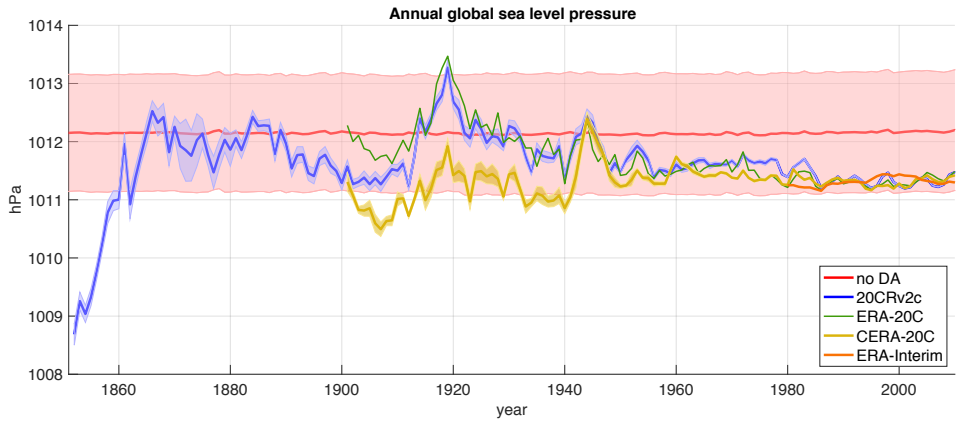

FIGURE 5 Time series of annual global SLP from several reanalyses, as well as a non-assimilating ensemble of model runs.

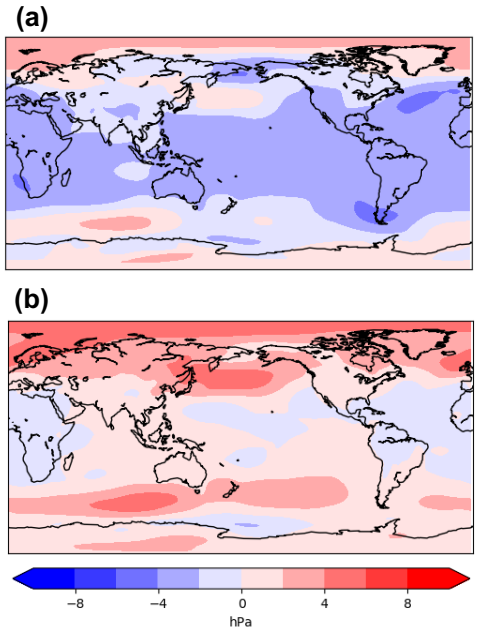

FIGURE 6 Analyzed sea level pressure anomalies (with respect to the analyzed 20CRv2c 1981–2010 climatology) for 1851–1853 of (a) 20CRv2c and (b) an identical experiment with the 20CRv2c system that assimilated 10% fewer ship observations.

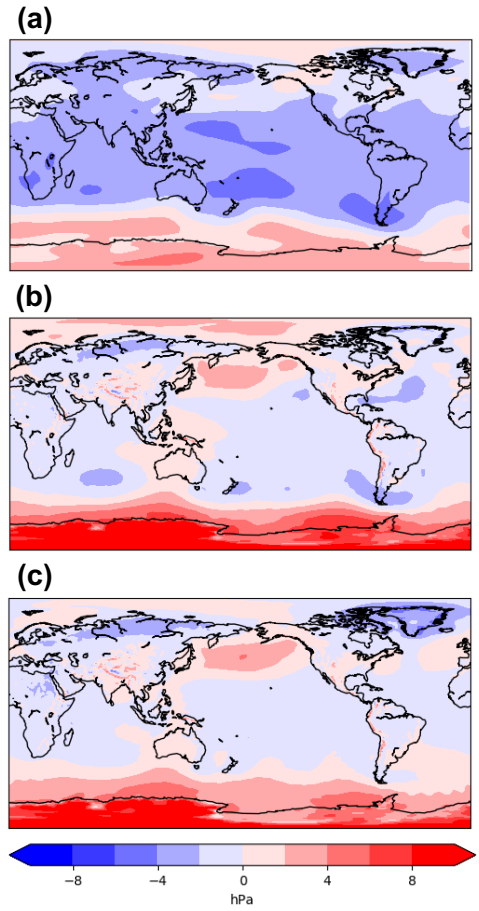

FIGURE 7 Annual-average sea level pressure anomaly (with respect to the 1981-2010 climatology) for 1854 of (a) 20CRv2c (without dry air mass specification), (b) preliminary version 3 tests without ship bias correction (with dry air mass specification), and (c) preliminary version 3 tests with ship bias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

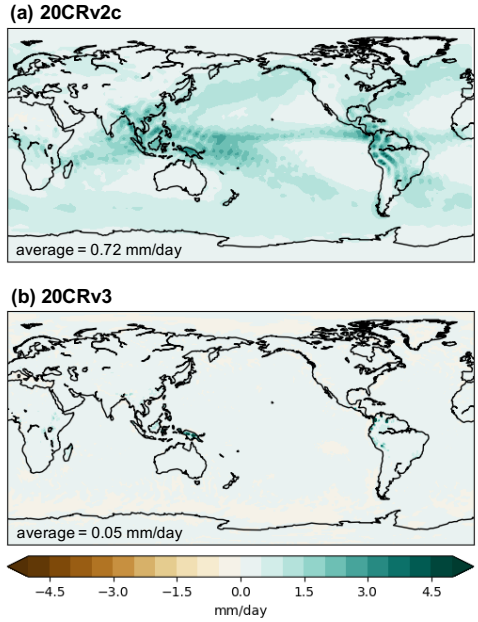

FIGURE 8 Annual average differences (mm day^{-1}) between 3-6 hour forecasted precipitation rate and 0-3 hour forecasted precipitation rate in (a) 20CRv2c and (b) 20CRv3 for 2002.

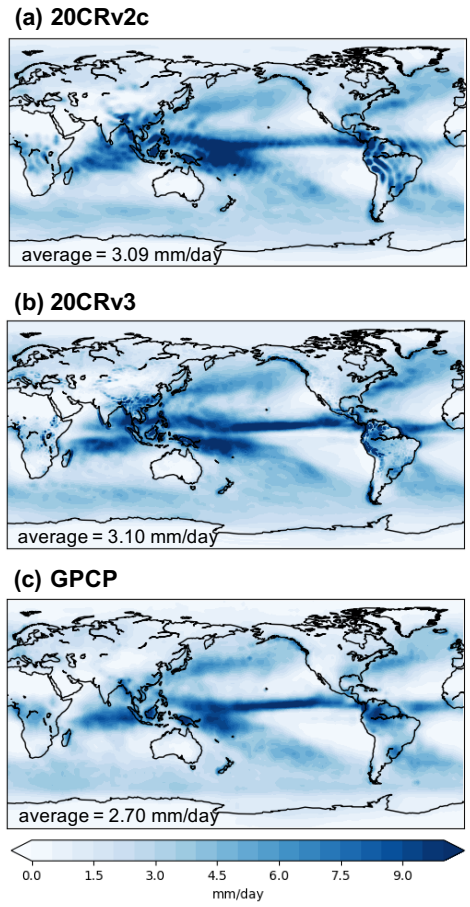

FIGURE 9 Annual average forecasted precipitation rate (mm day^{-1}) in (a) 20CRv2c, (b) 20CRv3, and (c) the GPCP satellite/gauge blended fields for 2002.

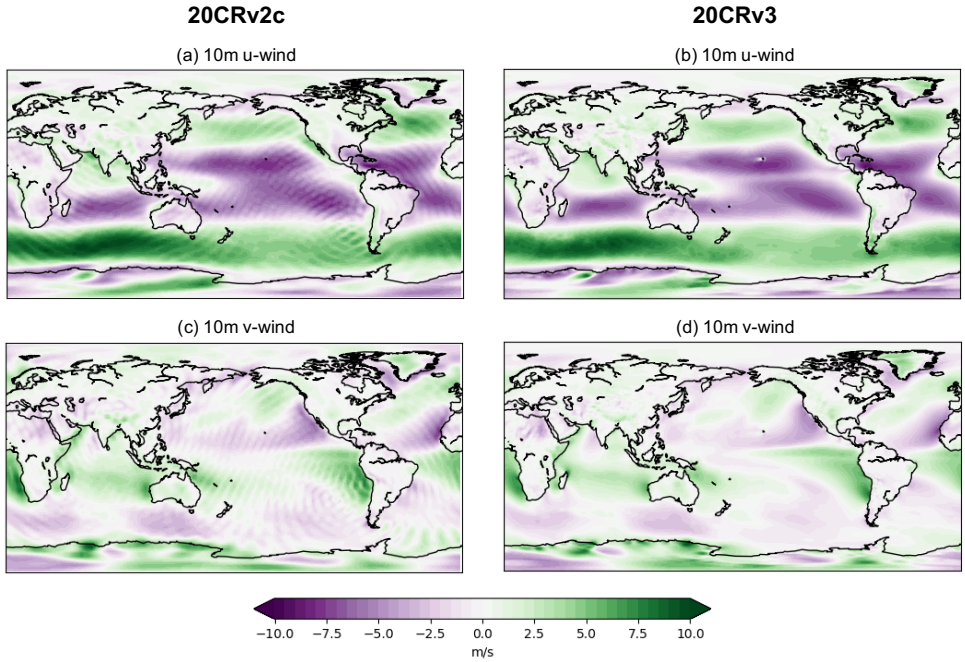

FIGURE 10 Annual average from 2002 of 10-meter zonal (top) and meridional (bottom) wind fields from 20CRv2c (left) and 20CRv3 (right), ms^{-1}.

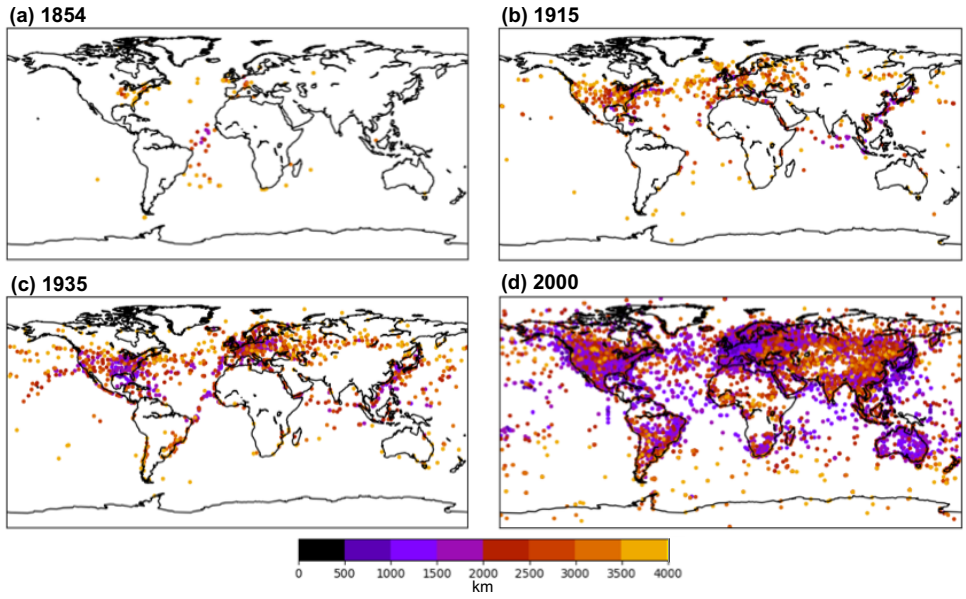

FIGURE 11 Localization values in km for all observations assimilated in 20CRv3 at 1200 UTC on 1 June for the year (a) 1854; (b) 1915; (c) 1935; and (d) 2000. Note that comparable plots for 20CRv2c would be entirely light orange (localization value of 4000 km).

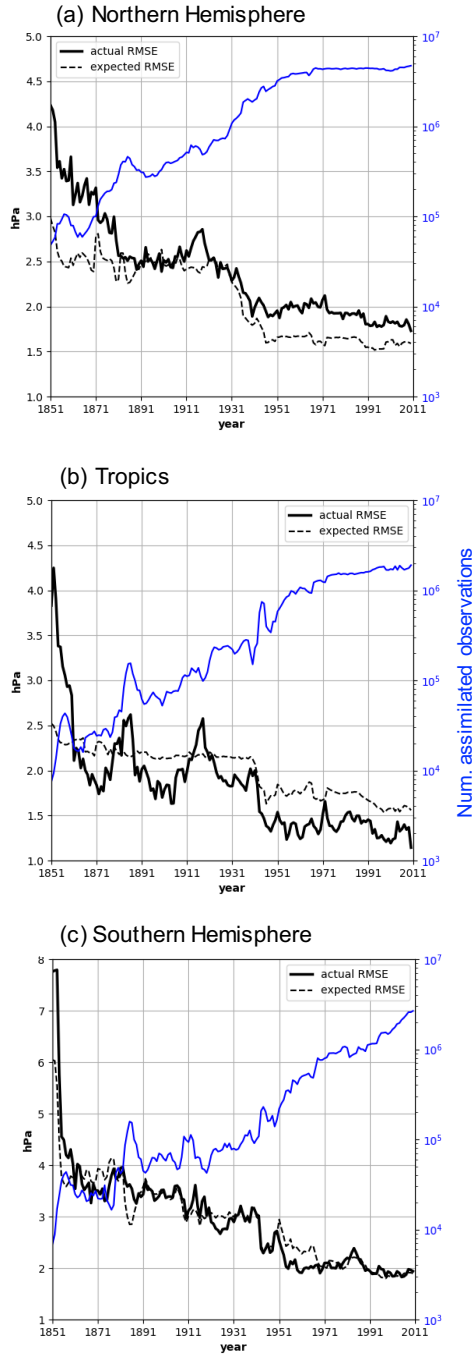

FIGURE 12 Actual (solid black) and expected (dashed) annual first-guess RMS errors for observations assimilated in 20CRv2c in the (a) the Northern Hemisphere (20°N to 90°N), (b) the tropics (20°S to 20°N), and (c) the Southern Hemisphere (90°S to 20°S). The number of observations assimilated in each region annually is plotted as a thin blue line (right-hand axis).

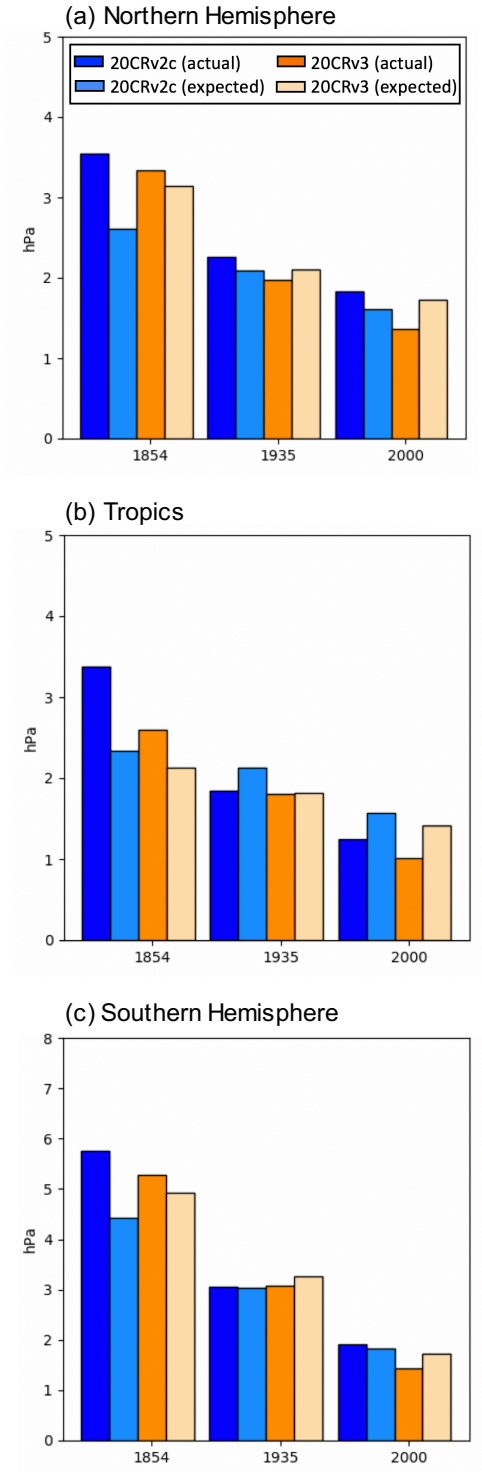

FIGURE 13 Actual and expected annual first-guess RMS errors for 20CRv2c (blue) and 20CRv3 (orange) for (a) the Northern Hemisphere (20°N to 90°N), (b) the tropics (20°S to 20°N), and (c) the Southern Hemisphere (90°S to 20°S) for three representative years.

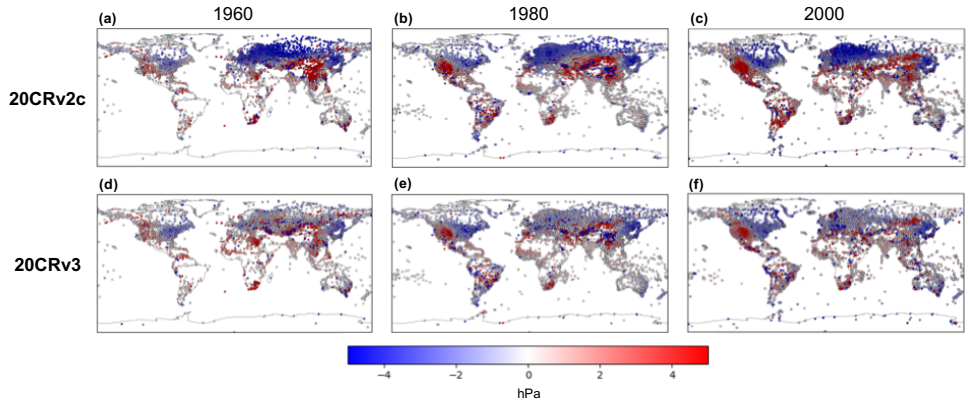

FIGURE 14 Annual average station pressure biases in hPa for 1960, 1980, and 2000, calculated from 20CRv2c (top row) and 20CRv3 (bottom row).

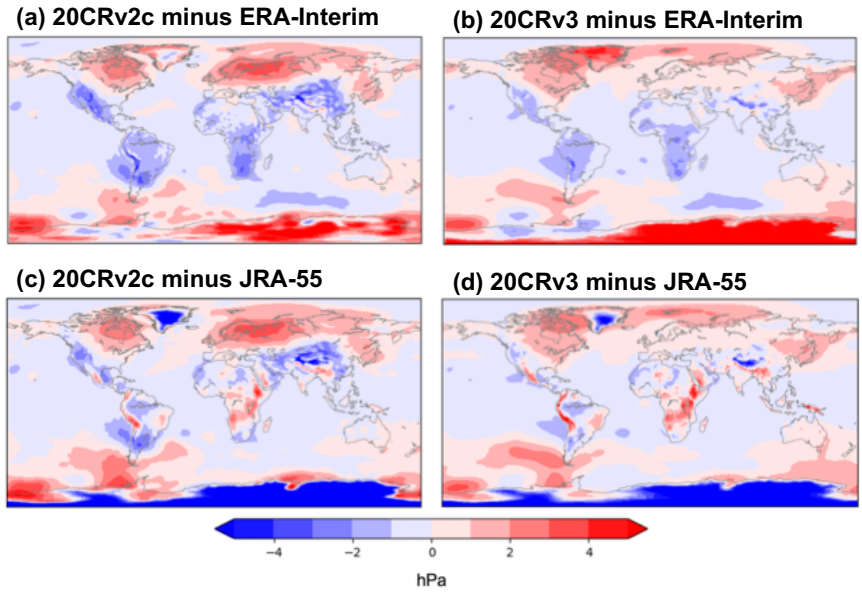

FIGURE 15 Annual average sea level pressure differences between the labelled for 2000 of (a) 20CRv2c and (b) 20CRv3.

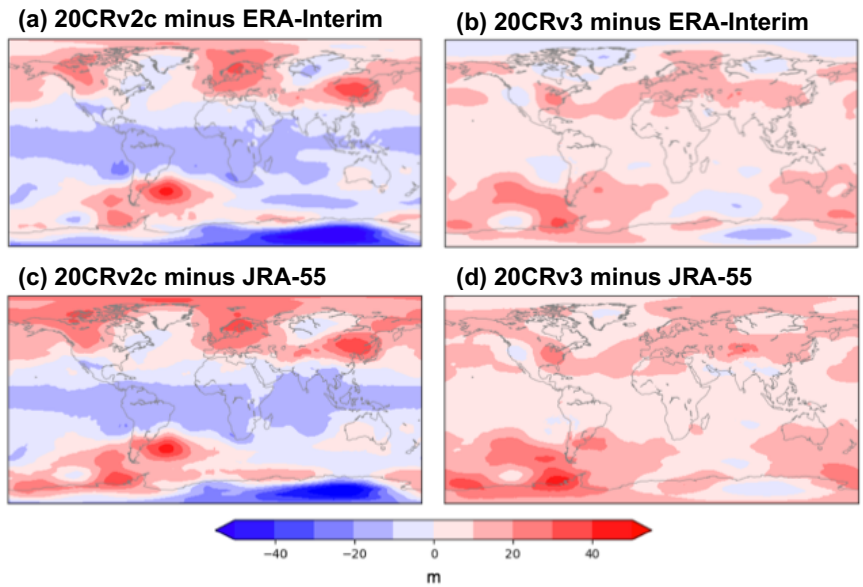

FIGURE 16 Annual average 500 hPa geopotential height differences from ERA-Interim for 2000 of (a) 20CRv2c and (b) 20CRv3.

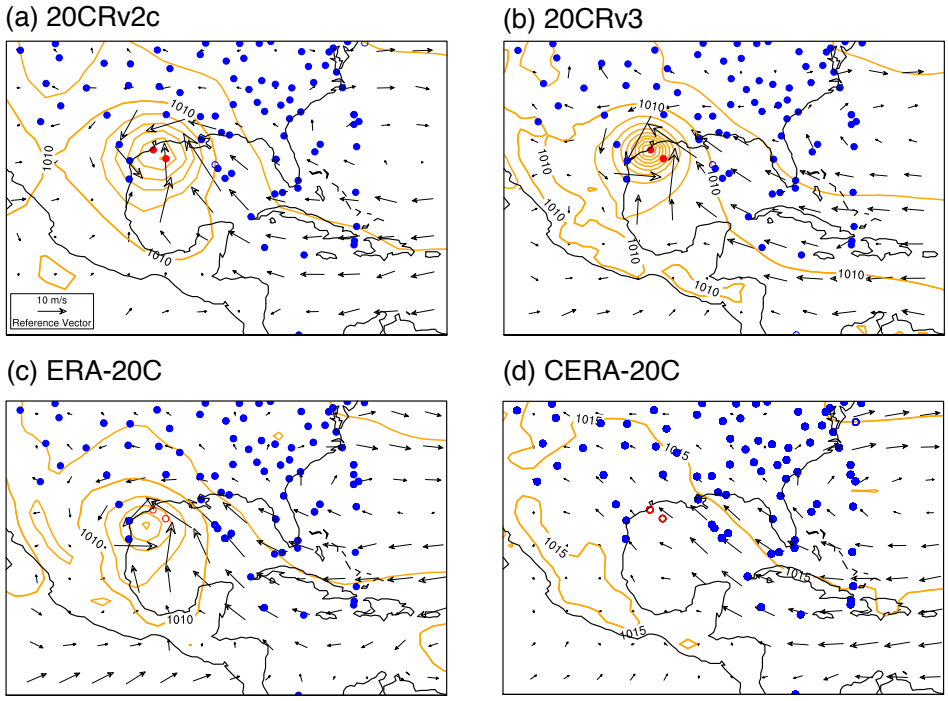

FIGURE 17 Sea level pressure (orange contours; interval by 5 hPa) and wind fields (vectors; ms^{-1}) for the 1915 Galveston hurricane, 17 Aug 1915 0600UTC, from (a) 20CRv2c, (b) 20CRv3, (c) ERA-20C, and (d) CERA-20C. Locations of available observations taken between 16 Aug 1915 2100 UTC and 17 Aug 1915 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.

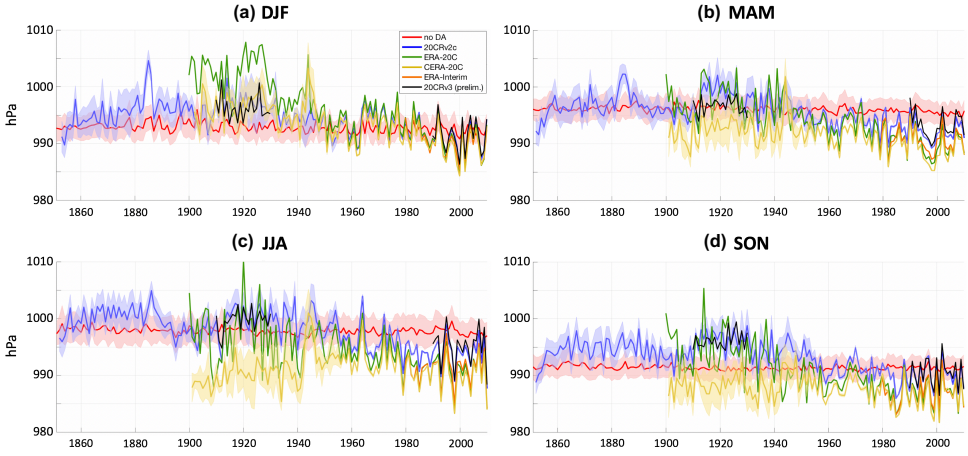

FIGURE 18 Time series of seasonal sea level pressure poleward of 60°S from 20CRv2c (blue), ERA-20C (green), CERA-20C (gold), ERA-Interim (orange), a non-assimilating model run using the 20CRv2c system (red), and preliminary 20CRv3 data without confidence intervals (black). Shading represents one standard deviation when available.