

# A gridded monthly upper-air data set from 1918 to 1957

Stefan Brönnimann · Thomas Griesser ·  
Alexander Stickler

Received: 27 July 2010 / Accepted: 24 October 2010 / Published online: 12 November 2010  
© Springer-Verlag 2010

**Abstract** Significant efforts have been devoted in recent years towards extending observation-based three-dimensional atmospheric data sets back in time. Such data sets form an important basis for a better understanding of the climate system. Here we present a new monthly three-dimensional global data set that is based on historical upper-air data and surface data. We use statistical reconstruction techniques, calibrated using ERA-40 data, to obtain gridded data from the numerous, but scattered and heterogeneous historical upper-air observations. In contrast to previous work, in which we used hemispheric principal components on both the predictor and the predictand side to reconstruct spatially complete fields back to 1880, we restrict the procedure to places and times where upper-air observations are available. Each grid column (consisting of four variables at six levels) is then reconstructed independently using only predictor variables in the vicinity (i.e., only local stationarity is required rather than stationary large-scale patterns). The product, termed REC2, is a gridded, global monthly data set of geopotential height, temperature, and  $u$  and  $v$  wind from 850 to 100 hPa back to 1918. The data set is sparse (i.e., many grid cells are empty), but provides an alternative to large-scale reconstructions as it allows for non-stationary teleconnections. We show the results of several validation experiments, compare our new data set with a number of existing data

sets, and demonstrate that it is suitable for analysing large-scale climate variability on interannual time-scales.

## 1 Introduction

There is a growing need for extending global climate data products backwards in time to have a better observational basis for analysing climate variability and comparison with model simulations. This is important for better understanding, assessing, and eventually predicting climate variability, and has long been done based on data from the Earth's surface. However, surface data alone do not provide an adequate depiction of atmospheric circulation throughout the whole depth of the troposphere, which is necessary for process-based analysis of climate variability.

Until recently, the time period accessible for studying atmospheric circulation at higher atmospheric levels than the surface was restricted to the periods since 1948 or 1957. These two dates mark the start of the NCEP/NCAR reanalysis (NNR, Kistler et al. 2001) and ERA-40 reanalysis (Uppala et al. 2005), respectively. The latter date is related to the International Geophysical Year (IGY), when a global radiosonde network was established, with standardised procedure, reporting, and instrument intercomparisons. Several radiosonde-based data products such as HadAT2 (Thorne et al. 2005), RAOBCORE (Haimberger 2007), RATPAC (Free et al. 2005), RICH (Haimberger et al. 2008) and the Iteratively Homogenised Radiosonde data set (Sherwood et al. 2008) also reach back to the IGY.

In recent years large efforts have been devoted towards extending the observation-based three-dimensional data sets of the atmosphere backward in time. Historical upper-air data have been compiled and digitised (Brönnimann

---

S. Brönnimann · T. Griesser · A. Stickler  
Institute for Atmospheric and Climate Science,  
ETH Zurich, Zurich, Switzerland

S. Brönnimann (✉)  
Oeschger Centre for Climate Change Research  
and Institute of Geography, University of Bern,  
Hallerstr. 12, 3012 Bern, Switzerland  
e-mail: broennimann@env.ethz.ch

2003; Grant et al. 2009a; Stickler et al. 2010) and statistical reconstructions of upper-level fields have been performed (Schmutz et al. 2001; Luterbacher et al. 2002; Brönnimann and Luterbacher 2004; Griesser et al. 2010; the latter is called REC1 hereafter). Recently, a 6-hourly three-dimensional data set resulting from an assimilation of surface pressure and sea-level pressure (SLP) data back to 1871 (the Twentieth Century Reanalysis or 20CR) has been released (Compo et al. 2010). These new data products are useful for studying large climatic anomalies such as the 1918 El Niño event (Giese et al. 2010), the early twentieth century Arctic warming (Grant et al. 2009b; Wood and Overland 2010), the “Dust Bowl” droughts (Brönnimann et al. 2009a; Cook et al. 2010), and the global climate anomalies during the 1940–1942 El Niño event (Brönnimann et al. 2004), or for statistical analyses of extreme circulation events, such as tropical cyclones (Emanuel 2010), and of large-scale circulation indices (Brönnimann et al. 2009b).

For many applications it may be advisable not to rely on one single data set as each has specific restrictions. Observations such as those from the Comprehensive Historical Upper-Air Network (CHUAN, Stickler et al. 2010) can often not be used “as is” because the data are not in gridded format, not on uniform levels (either altitude or pressure levels), and because observation times differ. Moreover, CHUAN data are not free of measurement errors, even though physics based adjustments have been applied to the radiosonde data, and the wind data have been quality flagged. The statistical reconstructions (REC1) have good statistical skill, but are based on the assumption of stationary large-scale patterns, which may be a disadvantage for some analyses. Finally, 20CR is not based on upper-level data and therefore may be less adequate for analyses at higher levels. For these reasons, we attempted to construct a data set that is as close to upper-air observations as possible but at the same time has the advantages of being on a uniform grid and exploiting information from the near neighbourhood to smooth out local errors.

Here we present a new, monthly reconstructed upper-air dataset derived from historical observations. Although a statistical reconstruction in principle, the data set is close to the historical data by construction while alleviating the stationarity assumption as far as possible. As a consequence, the new data set is suitable for analysing large-scale climate variability on interannual time scales, including detection of non-stationary teleconnections.

In this paper we summarise the data used, outline the reconstruction and validation techniques and show comparisons with other data sets. Finally, we demonstrate the potential of the new product using two examples of prominent climatic variations in the reconstruction period, namely the 1939–1942 El Niño (perhaps the strongest

large-scale climate signal on interannual time scales that can be found in the twentieth century) and the Dust Bowl droughts of the 1930s.

## 2 Data and methods

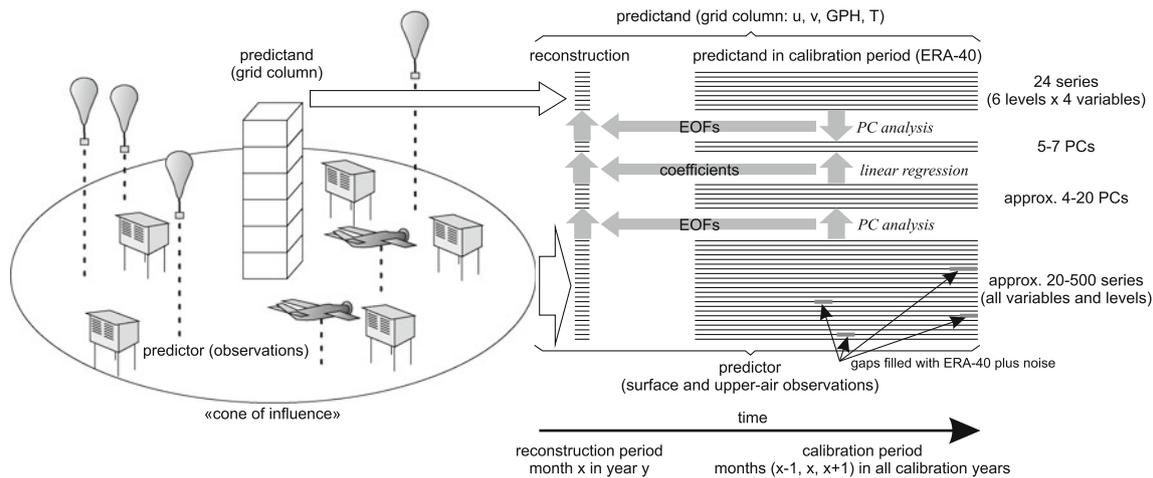
### 2.1 Overview of the procedure

The reconstruction approach requires various types of data sets and involves many steps. Here we first provide a brief overview of the procedure and the terminology used. It is also schematically shown in Fig. 1; Table 1 summarizes the main assumptions and data sets used and compares these features with REC1, i.e., the approach of Griesser et al. (2010). Principal component (PC) regression relies on establishing a statistical relation between predictor data and predictand data in several steps (namely PC analysis and regression). Note that prediction in this case is not a forecast in time based on the current state, but rather an estimation based on (spatially) neighbouring information. The predictand data are the dependent variables in the statistical relation, the predictor data are the independent variables. Establishing the statistical relation is termed calibration and implies that both predictor data and predictand data need to be available for a common period, termed calibration period. Once the relation is established, it can be applied to past predictor data (the statistical relation in this context is then often called transfer function) to obtain an estimation of the predictand in the past. This estimation is termed reconstruction.

### 2.2 Data sets used for reconstruction

The main predictor data set is formed by historical upper-air data from the CHUAN data set (Stickler et al. 2010). This data set comprises 12.6 million profiles prior to 1958; the radiosonde stations are then supplemented with data from the Integrated Global Radiosonde Archive (Durre et al. 2006) afterwards, using the RAOBCOREv1.4 correction (Haimberger 2007). We use the data in the form of monthly means after adjustment of the data for known or suspected physical errors (such as uncorrected radiation and lag errors) or adjustment for statistically detected inhomogeneities (version CDCRM, see also Brönnimann 2003; Ewen et al. 2008a; Grant et al. 2009a; Stickler et al. 2010). Note that we have kept series classified by Stickler et al. (2010) as “suspicious” (this could be, e.g., wind series with a bias, with respect to a reference series, in one of the two wind components between the 90 and the 95% significant level).

In the reconstruction process, the upper-level predictors are supplemented with surface predictors. We used sea-level



**Fig. 1** Schematic depiction of the reconstruction approach. *Left* spatial depiction of predictor and predictand data and the concept of a “cone of influence”. *Right* reconstruction approach

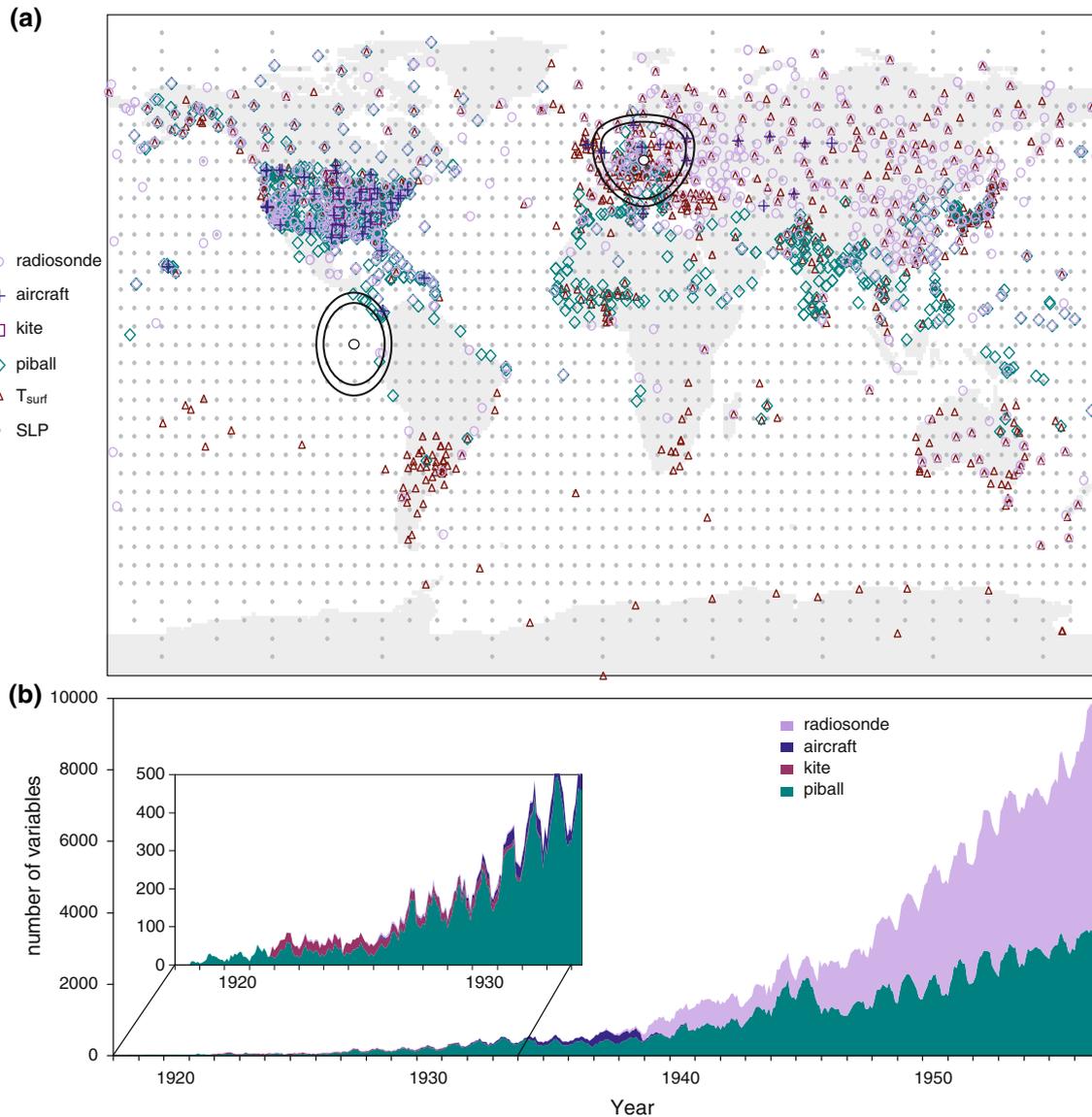
**Table 1** Comparison for the assumptions and procedures in REC1 and REC2

Approach	PC regression	
	REC1	REC2
Reconstructed variables	GPH, <i>T</i> at six levels (850, 700, 500, 300, 200, 100 hPa) <i>u</i> , <i>v</i> at 3 km	GPH, <i>T</i> , <i>u</i> , <i>v</i> at six levels (850, 700, 500, 300, 200, 100 hPa)
Time period	1880–1957	1918–1957 (calibrated data using the statistical model for 1957 are given for 1958–2000)
Spatial scale and resolution	Global 2.5° × 2.5° (regridded from an equal area grid), spatially complete	Global 2.5° × 2.5°, incomplete
Subdivision in space	Three regions (NH extratropics, tropics, SH extratropics)	Grid column by grid column
Stationarity assumption	Hemispheric spatial patterns	Vertical grid columns
Predictor in reconstruction period	Station SAT, gridded SLP, upper-air observations (GPH, <i>T</i> , <i>u</i> , <i>v</i> ; pibal winds were supplemented in 1948–1957 with NNR winds)	Station SAT, gridded SLP, upper-air observations (GPH, <i>T</i> , <i>u</i> , <i>v</i> )
Predictor in calibration period	Station SAT (gaps filled with ERA-40), gridded SLP; all upper-air series from ERA-40	Station SAT (gaps filled with ERA-40), gridded SLP; upper-air series from observation (gaps filled with ERA-40)
Predictand in calibration period	ERA-40	ERA-40

pressure (SLP) data from HadSLP2 on a 5° by 5° grid (HadSLP, Allan and Ansell 2006), with fewer cells at higher latitudes similar to a latitude weight (see Brönnimann and Luterbacher 2004, hereafter BL2004; see also Fig. 2). Moreover, surface air temperature data from the NASA-GISS station data set were used. We selected stations with good temporal coverage (even complete temporal coverage was required for the USA, where the number of stations is high) and with a high correlation with ERA-40 reanalysis data (which assures that these stations are representative for a larger spatial scale). In total 760 stations were selected (Hansen et al. 1999; see Griesser et al. 2010 for details). Missing values in these surface data after 1957 were filled as

in BL2004 after the standardization procedure (see below) by means of standardised anomalies of 925 hPa temperature in ERA-40.

As predictand data set for the calibration period we chose ERA-40 (Uppala et al. 2005). Errors in that data set are documented (e.g., Bengtsson et al. 2004; Bromwich and Wang 2005; Grant et al. 2008), but the only other alternative, NNR, also suffers from problems that may affect, in particular, the low-frequency variability (see e.g., Trenberth et al. 2001; Harnik and Chang 2003; Santer et al. 2004; Grant et al. 2009a). In many aspects ERA-40 is superior to NNR reanalysis (see e.g., Simmons et al. 2004). Experience from varying the calibration period in previous work showed that



**Fig. 2** Location of predictors used in this study (a). The *distorted circles* show the cones of influence around two grid points [15°E, 50°N] and [90°W, 0°N], marked in *white*. Number of upper-air monthly mean values as a function of time (b)

30–40 years of data are needed to obtain robust reconstructions. Consequently, other reanalysis data sets are too short.

### 2.3 Data sets used for validation

The reconstructions are compared with other historical upper-air data products. In addition to comparisons with the CHUAN data set itself, we also use other upper-level reconstructions (Schmutz et al. 2001, BL2004; Griesser et al. 2010) as well as 20CR (Compo et al. 2010). Note that all reconstructions are partially based on the same predictor data sets and hence are not fully independent. There is also some dependence between this reconstruction and 20CR as

both use SLP data (although 20CR used them on a subdaily scale and the reconstruction on a monthly scale).

Fully independent comparisons can be performed with historical total ozone data (Brönnimann et al. 2003, see also Griesser et al. 2010), which at midlatitudes are expected to be strongly correlated with temperature and geopotential height in the upper troposphere and lower stratosphere. Upper-level troughs are accompanied by a low tropopause and therefore high total ozone and high temperatures in the lower stratosphere, but relatively low temperatures (at the same pressure levels) in the upper-troposphere and low pressure. The opposite is the case for upper ridges. The relation is not restricted to synoptic time

scales, but still appears on the monthly and interannual time scale. Hence monthly total ozone is expected to be negatively correlated with temperature in the upper-troposphere but positively in the lower stratosphere. For GPH we expect a negative correlation at all altitudes, which maximizes in the lower stratosphere (Brönnimann et al. 2000). We used monthly mean values of total ozone at Arosa (Switzerland), 1926–2002 (Staehelin et al. 1998), Tromsø (Norway), 1935–1972 (Hansen and Svenøe 2005), Oxford (UK), 1924–1975 (Vogler et al. 2007), New York (USA), 1941–1944, and Zi-Ka-Wei (China), 1932–1942 (both from Brönnimann et al. 2003). Where no station data were available after 1978, TOMS total ozone data (Version 8) were used to supplement the series.

#### 2.4 Data sets used for analysis

We also compared the reconstructions with results from AMIP-type climate model simulations covering the twentieth century. In these simulations, all forcings are prescribed as realistically as possible, including sea-surface temperatures (SSTs). Reconstructions and simulation results are fully independent as they do not use common data or common assumptions.

We used two sets of ensemble simulations of the twentieth century, namely an ensemble of six simulations performed with the HadAM3 model (EMULATE runs) as well as an ensemble of nine simulations performed with the chemistry-climate model SOCOL (Fischer et al. 2008b). These simulations are also compared and discussed in Scaife et al. (2009). SOCOL was chosen here because a specific target of previous work was the 1940–1942 climate anomalies (which included strong anomalies in total ozone, see Fischer et al. 2008a), which will be further discussed in Sect. 4.

#### 2.5 Preparation of predictand data

The goal of this study is to reconstruct monthly upper-level geopotential height (GPH), temperature, and wind ( $u$  and  $v$  components) for the global atmosphere on a  $2.5^\circ$  by  $2.5^\circ$  grid (see also Fig. 1). We chose to reconstruct the following levels: 850, 700, 500, 300, 200, and 100 hPa. The 1,000 hPa level was not reconstructed because systematic differences between the predictors (i.e., observations) and our calibration data set (i.e., ERA-40) in the planetary boundary layer might lead to errors.

As a first step of the reconstruction procedure, the corresponding levels and variables were extracted from ERA-40 and each grid column of the  $2.5^\circ$  by  $2.5^\circ$  grid (containing four variables at six level, see above and Fig. 1) was stored in a separate file. All data series were then standardised based on the period 1958–2001.

#### 2.6 Preparation of predictor data

The first step in the preparation of the predictor data is to subsample the vertical resolution. Rather than to keep all 83 levels reported in CHUAN, we only used the levels 1, 2, 3, 4, 5, 6, 8, 10, 15 km in the case of altitude level data and 850, 700, 600, 500, 400, 300, 250, 200, 100 hPa in the case of pressure level data. Because a large number of stations started observations during 1957 in preparation for the IGY, we “froze” the network in December 1956 to avoid problems due to oversampling.

Figure 2a shows a map of the predictor data used. Note that although there are many stations, most of them have data only over a limited time span. Figure 2b shows a time series of the number of available upper-level monthly mean values, indicating that only 5–10 stations are available for the first few years (reporting only few levels), then the number gradually increases. In the first years, pilot balloon data dominate. Kite and aircraft data (mostly temperature and geopotential height, or just temperature) become somewhat more abundant in the 1920s and 1930s, but still comprise only a minor fraction. The advent of radiosonde instruments in the 1930s changes this only slowly and it is not until the period after the Second World War that the radiosonde became the dominating platform. Concerning the spatial coverage, the northern midlatitudes are well covered, as expected. Somewhat less expected might be the good coverage of the Arctic, while the tropical regions and especially the southern extratropics are very sparsely sampled. Obviously, ocean regions are poorly sampled.

A complete data set is required for calibration after 1957, but some series end earlier and many have gaps. For instance, kite soundings were no longer performed after the 1930s. Therefore, we filled all gaps in the calibration period with corresponding ERA-40 data, interpolated to the location and level of the historical data. Note that this contrasts with the approach in Griesser et al. (2010), where the predictor data in the calibration period were taken entirely from ERA-40 (not just the gaps).

The interpolated reanalysis data are expected to exhibit less variability than the observational data as they do not represent local features and smooth out random errors and to some extent biases. This creates the danger of overfitting as there is too little noise in the calibration period. To account for this, we perturbed the interpolated reanalysis data (i.e., all filled gaps) by a stochastic error model mimicking the differences between observations and reanalysis in a realistic way (see Appendix).

#### 2.7 Principle component regression

The basis of our reconstruction is principal component regression, which has been used in many previous studies

(Jones et al. 1987; Cook et al. 1994; Luterbacher et al. 2002, 2004, BL2004). When applying this method to large spatial fields, a high skill may result because of dominating patterns such as El Niño/Southern Oscillation (ENSO) and its teleconnections. However, the underlying assumption is that these large-scale patterns (i.e., the teleconnections) are stationary, which may not be the case. To avoid this problem, we use the much weaker assumption of local stationarity. Each grid column is reconstructed as a whole (temperature, GPH,  $u$  and  $v$  wind at six levels, see Fig. 1), but independently of all other grid columns, using only predictor data from the vicinity (i.e., a “cone of influence”, see Fig. 1). In this way it is assured that opposite centres of action do not influence each other (e.g., the Azores high does not constrain the Icelandic low).

We chose a maximum distance that approximately represents a spatial correlation of  $>0.5$ , which corresponds to 1,500 km for upper-level temperature and GPH and 1,200 km for upper-level winds and surface variables (see Griesser et al. 2010, for more details). This “cone of influence” is shown in Fig. 2a for a grid point in Europe (50°N, 15°E) and one near the Galapagos Islands (0°N, 90°W). It can be seen that in the former case, a large number of stations is available in the vicinity of the grid point. In the second case, there is but one pilot balloon station and two radiosonde stations, all of which are in mainland Ecuador and hence relatively far away (the pilot balloon and radiosonde stations in central America are outside the 1,200 and 1500 km limits, respectively).

The actual reconstruction approach then proceeds grid column by grid column, month by month (see Fig. 1). In other words, our “reconstruction” in effect is composed of hundreds of thousands of independent reconstruction. For a given month and grid column, the predictors available on that month in the vicinity of that grid column are identified and are selected in the calibration period. To reconstruct a grid column for a given month, there must be at least three upper-level variables in the cone of influence, i.e., a monthly mean temperature profile from aircraft or kites with values at 1, 2 and 3 km or 850, 700, and 600 hPa would suffice as a minimum.

Having identified the predictors, a 3-calendar months moving window was chosen for calibration, i.e., for reconstructing May 1922 we used only data from April, May and June in the calibration period. Surface and upper-level predictor variables are then pooled together and weighting was only performed if the surface contributed to less than 25% or more than 50% (in which cases the weight was set to these limits) to assure that vertical information is adequately considered.

The next step is a reduction of variables on both sides (predictor and predictand) by a principal component analysis. Statistical criteria could be used to find the best

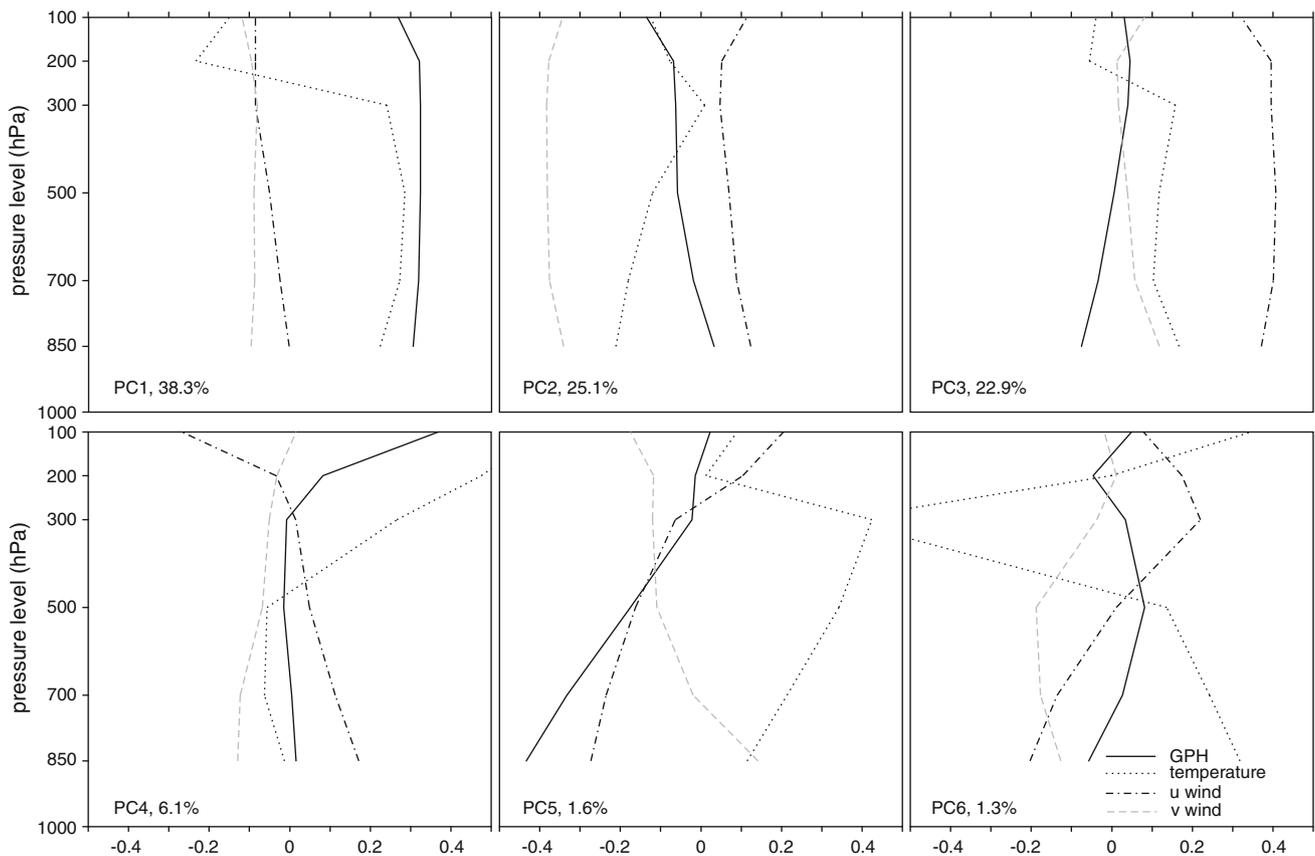
truncation in each case, but this would make calibration and validation incomparable. Another approach is to optimise the fraction of retained variance simultaneously at both the predictor and the predictand side based on validation experiments. We used this approach in previous work (Griesser et al. 2010), but the danger of overfitting exists. In our present case, in which we compose a data set from countless individual reconstructions, we prefer a simple approach and use fixed thresholds based on experience from previous work (Schmutz et al. 2001, BL2004, Griesser et al. 2010).

On the predictand side (in the calibration period, i.e. in the ERA-40 grid columns), the retained variance was fixed to 95%. Usually 5–7 PCs are required to explain 95% of the variance in all 24 series (four variables at six levels). The empirical orthogonal functions (EOFs) or loading patterns pertaining to these PCs have a rather simple structure and can often be interpreted physically. Figure 3 shows the selected EOFs for the grid point 50°N, 15°E for January. The numbers in the panels indicate the variance explained by this EOF. The first EOF has large loadings in GPH at all levels. It captures high or low pressure. Temperature shows a corresponding pattern, with a sign change at the tropopause. The second and third EOFs, which explain a very similar amount of variance, have large loadings in  $v$  and  $u$  wind and hence represent advective regimes (meridional and zonal advection, respectively). The remaining EOFs explain much less variance. The fourth EOF is a stratospheric type and represents lower stratospheric warmings and coolings. The fifth and sixth EOF appear to be mixed or shear flow types.

These EOFs are assumed to be stationary, which in reality may not be the case. However, we think that this assumption is better justifiable than stationarity of large-scale 3D patterns.

On the predictor side, we set the fraction of explained variance to 90% as we assume that these data contain more non-instrumental noise. As for the predictand, we expect the predictor PCs to capture more simple regimes which may have a physical meaning. This includes monopoles, gradients, or regional dipoles that may be related to land surface properties and orography (e.g., gradient across the Alps), but not “centres of action” of teleconnections. Due to the heterogeneity of the predictor data (different vertical coordinates, variables, etc.), this assertion is more difficult to confirm and was only sporadically checked.

Each predictand PC time series was then expressed as a linear combination of the selected subset of predictor PC time series using linear regression. We used a least-squares estimator because the number of regression models to be solved is extremely large (millions). Robust methods could lead to a slightly better skill, but at the price of computational efficiency.



**Fig. 3** Loading patterns of the first six predictand PCs for 50°N/15°E in January (unit is standard deviations). The numbers in the *panels* indicate the variance explained by the corresponding PC

The predictand PCs were then calculated from the predictor PCs in the reconstruction period and from the regression coefficients. The reconstructed anomaly field is a linear combination of the reconstructed predictand PCs multiplied by the predictand PC scores from the calibration period. Finally, the standardization procedure was reversed and the subtracted climatology added.

The reconstruction was performed up to 1957. After 1957, we applied the last valid calibrated model for each calendar month (mostly from the year 1957) until 2000. Comparing this with other data sets defines an upper-limit of the quality of our reconstructions. This is because the number of predictor variables is usually high in the last valid model, because some of the predictor data may actually be from ERA-40, and because it is not independently calibrated.

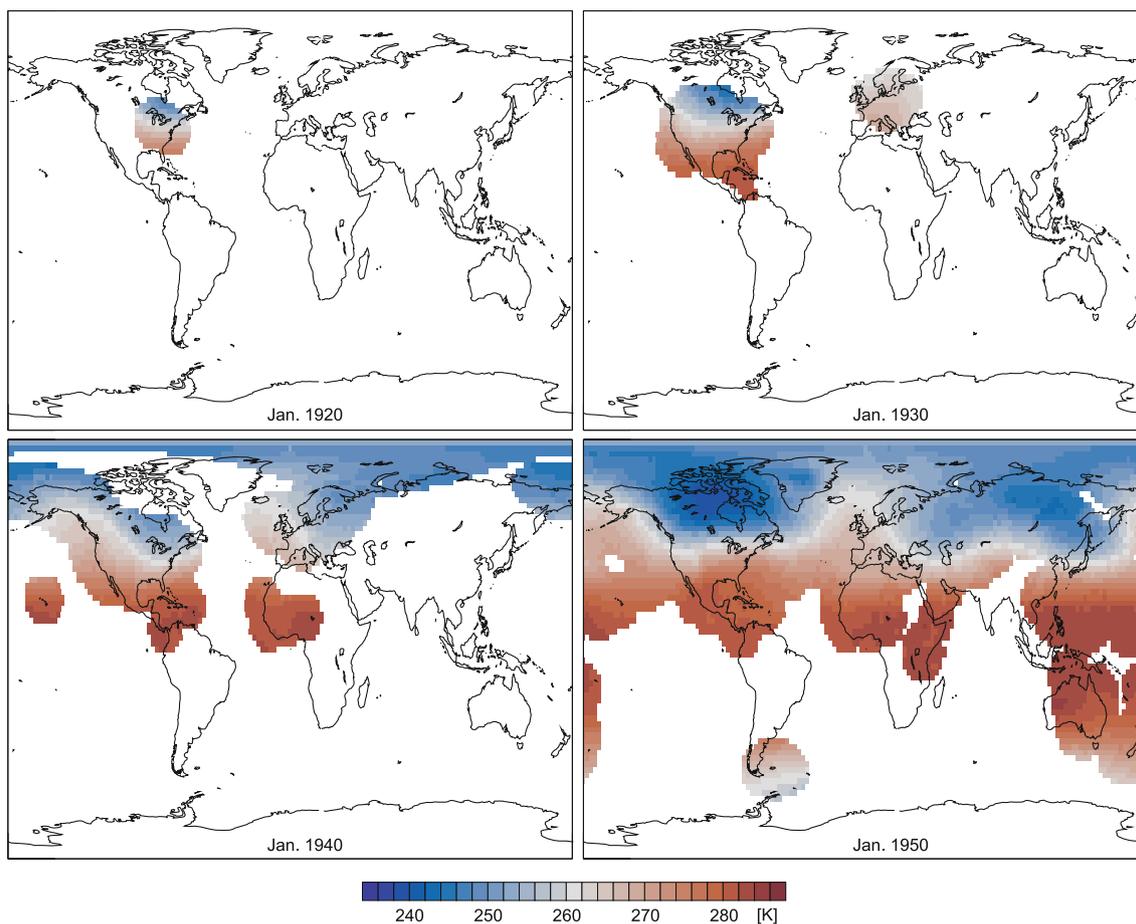
## 2.8 Validation

The skill of the reconstruction approach and the error of the reconstructions was tested in several ways: split-sample validation, comparison with other (partly dependent) data products, and validation with independent data (total

ozone). In previous work (Ewen et al. 2008b) we also validated the approach in a surrogate climate (i.e., the control run of the CCSM3.0 coupled climate model). The conclusion was that the split-sample validation usually gives a good estimation of the true skill of the reconstruction and that the calibration period is long enough to give robust reconstruction. However, validation in a climate model might be too optimistic as errors or inadequacies in the predictand data cannot be accounted for (e.g., a misrepresentation of the boundary layer or of topography) and results might depend too strongly on the stochastic error model used in the predictor data. Hence, we rather try to compare the final reconstructions with other data products as much as possible.

In the split-sample validation we used only part of the sample for calibration and the other part for validation. In this case, we performed two experiments in which we retained either the first or last third of the data for validation (calibrate in 1972–2001 and validate in 1958–1971 or calibrate in 1958–1987 and validate in 1988–2001), as in BL2004 and Griesser et al. (2010).

Several measures were used for quantifying the reconstruction uncertainty: averaged bias, root mean squared



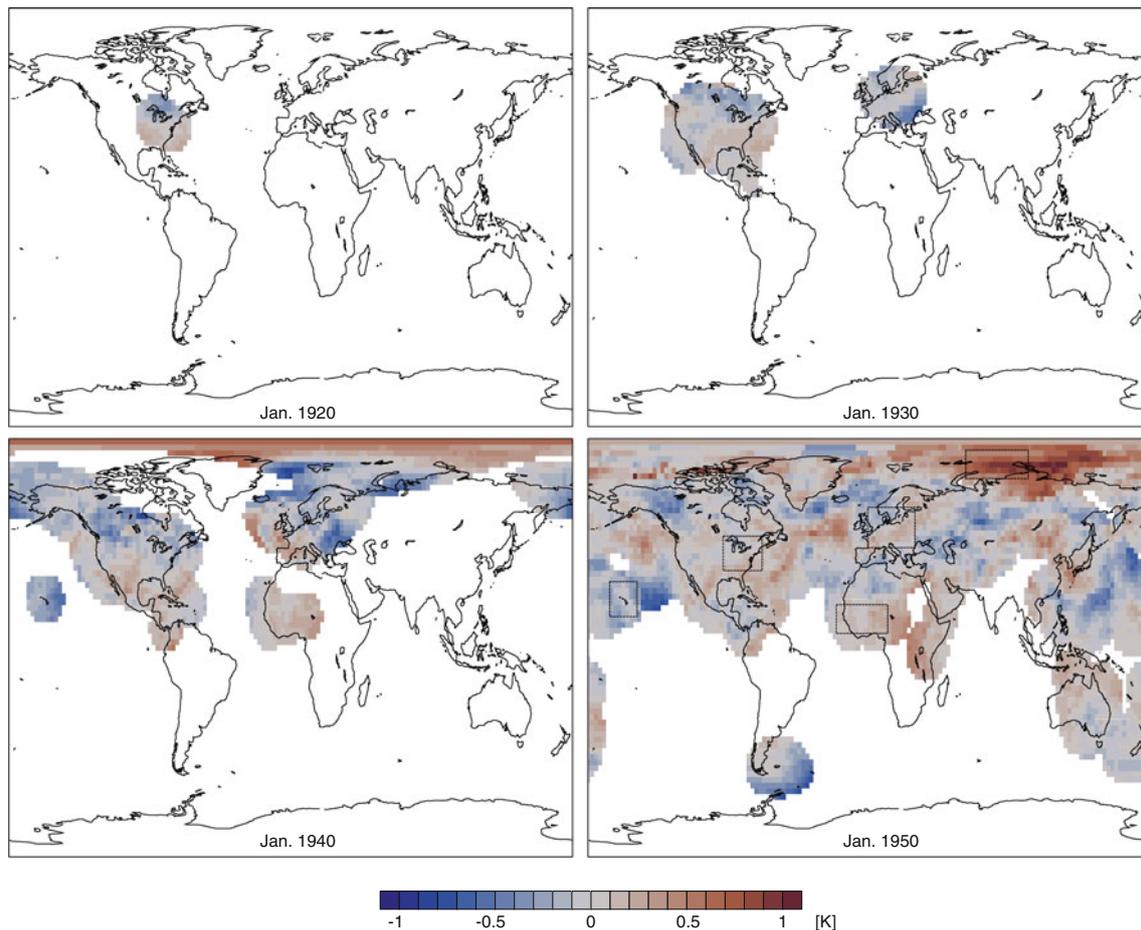
**Fig. 4** Reconstructed temperature at 700 hPa in January 1920, January 1930, January 1940, and January 1950, respectively

error (RMSE), and reduction of error (RE, see Cook et al. 1994). Values of RE can be between  $-\infty$  and 1 (perfect reconstruction). An RE of 0 is indicative of a reconstruction not better than a no-knowledge prediction (in our case climatology), whereas an  $RE > 0$  points to a model with predictive skill.

### 3 Results

The reconstruction is very sparse in the beginning. The first month with any data is October 1918. Figure 4 shows as an example the reconstructions of 700 hPa temperature for January 1920, January 1930, January 1940, and January 1950. In the first years the reconstruction is restricted to the eastern USA, later it encompasses more of North America as well as Europe. During the 1930s the coverage increases significantly and from January 1940 on includes Western Africa. In 1950, coverage is more than 50% of the globe, but the tropical oceans and the Southern Hemisphere are still very poorly covered.

The error measures pertaining to these fields are shown in Figs. 5, 6, 7. The interpretation of the bias error (Fig. 5) is not straight forward as it refers to the mean error of a given model (i.e., a combination of variables) over the two validation periods. This error does not necessarily appear as a time-invariant bias in the reconstruction (i.e., over many models), as can be seen by comparing the bias in the four different cases (Fig. 5). Although the predictand data (calibration and validation) is the same in all four cases, the bias differs because it is a function of the combination of available predictors rather than of the calibration data. If, on the other hand, the bias is relatively constant over time, this can point to systematic errors in the approach (e.g., strongly non-gaussian residuals). In our examples shown in Fig. 5, the bias shows no clear large-scale pattern but exhibits a clear maximum in the Arctic, which might point to an inadequacy of the approach in that region. We analyse the skill and error measures for this region in more detail in the following paragraphs. Some of the features in the bias fields are consistent over time (southern Ukraine or the Great Lakes region).



**Fig. 5** Bias of the reconstructed 700 hPa temperature in January 1920, January 1930, January 1940, and January 1950, respectively. The *dashed areas* indicate the regions for which error measures are averaged and displayed in Figs. 7 and 8

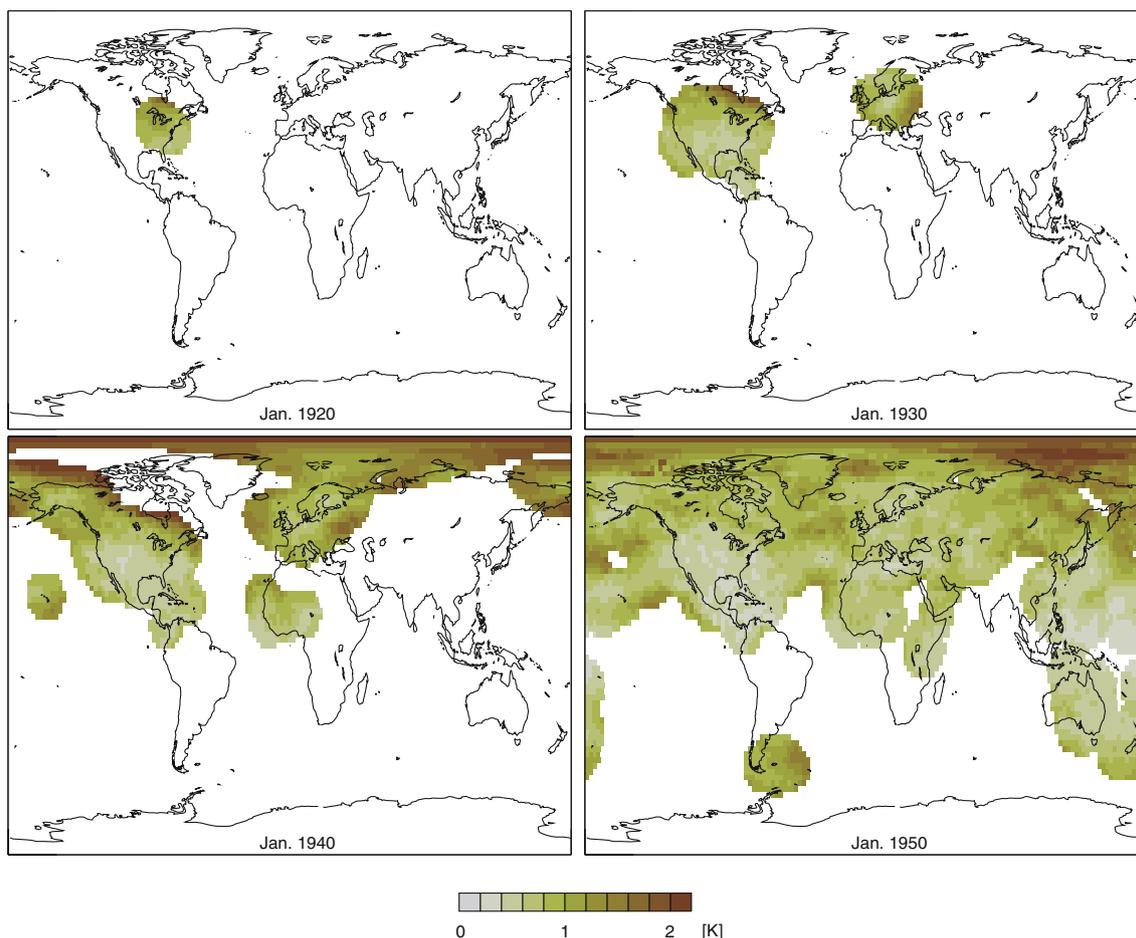
The root mean squared error (Fig. 6) is arguably the most widely used measure of the uncertainty of a reconstruction in units of the variable and can be used to estimate confidence intervals. In this case, it reflects both the skill of the reconstruction and the variance of the variable. As a consequence, the RMSE of temperature is larger in the polar regions than in the tropics. At midlatitudes, although the variance is large, the RMSE is small due to the high skill of the reconstruction. The RMSE shows peripheral effects, i.e., larger error at the edges of the reconstruction region.

Finally, Fig. 7 shows the RE statistics, which is a dimensionless measure of the skill of the reconstruction and thus allows comparison across levels, variables, and regions. Since we reconstruct only grid columns with sufficient local predictor data, we expect high skill everywhere and RE is larger than 0.5 in a very large fraction of the area. However, there are notable exceptions including (apart from the peripheral areas) the tropical monsoon regions and the Russian Arctic. The low skill in the West African monsoon region could be due to a bad

representation of the monsoon in reanalysis data sets with a seasonally varying bias (Stickler and Brönnimann, submitted manuscript). Even if the observations are good, they are calibrated against inadequate data and hence reconstructions will likely not be good. Furthermore, in this situation the gap-filling on the predictor side tends to make the skill even worse. We will discuss this region in more detail below.

In order to address the error measures also for other levels and variables, we computed area-averaged RE and RMSE for five areas (shown in Fig. 5), eastern North America [92.5°W–70°W, 35°N–47.5°N]; Central Europe [0°E–27.5°E, 45°N–60°N], the Russian Arctic [62.5°E–100°E, 75°N–85°N], the subtropical North Pacific [165°W–150°W, 15°N–27.5°N], and tropical West Africa [17.5°W–10°E, 7.5°N–17.5°N]. Together, these regions span a large range of the characteristics discussed above and include two areas with generally good skill and three areas with low or even very low skill.

Figures 8 and 9 show the time series of RMSE and RE for these regions. Noteworthy are the very high RMSE



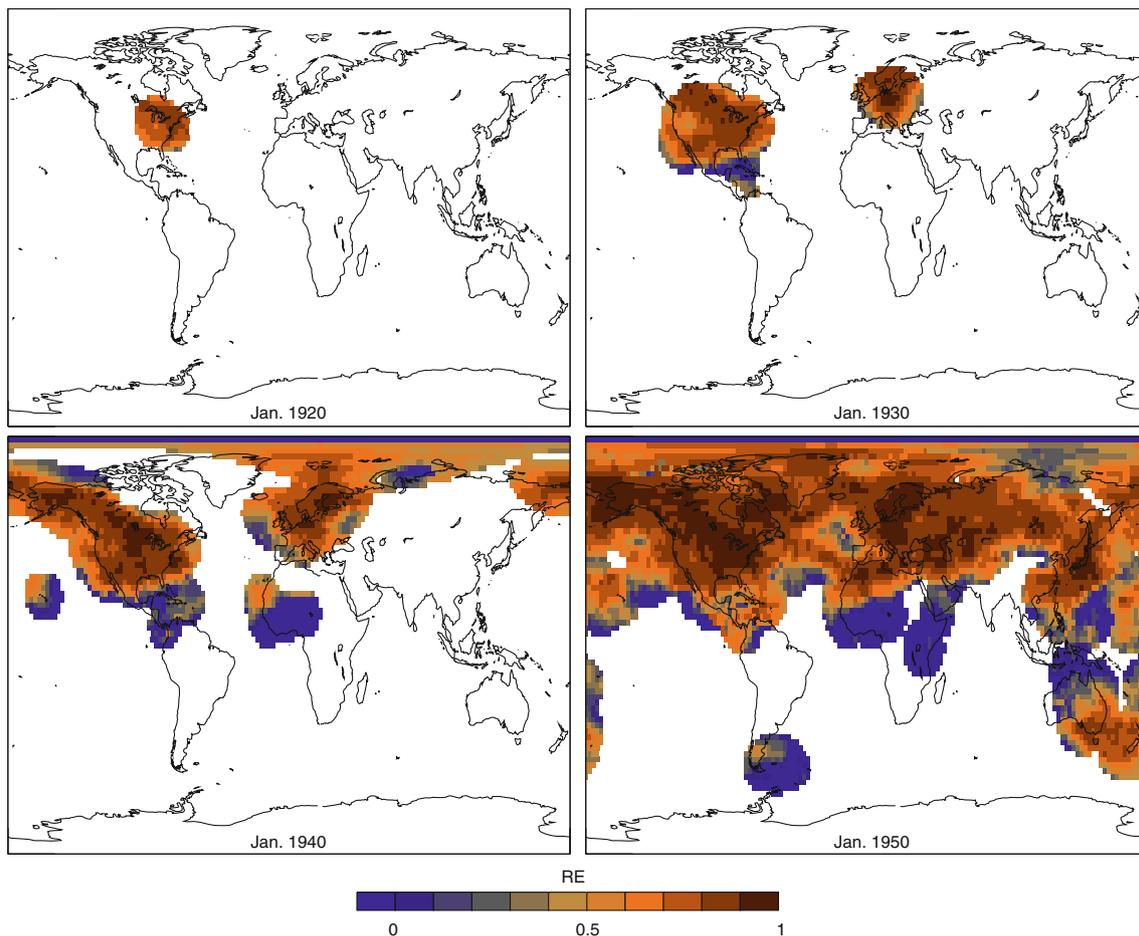
**Fig. 6** RMSE of the reconstructed temperature at 700 hPa in January 1920, January 1930, January 1940, and January 1950, respectively

(Fig. 8) for Siberia for all levels and variables. Another general feature is the large annual cycle of RMSE and RE for all variables, levels, and regions. Within each region, there is a tendency towards a smaller RMSE with time. Thus, the addition of predictor variables (see Fig. 2) not only increases the coverage but also decreases the error of the reconstructions.

As rule of thumb, we find that the skill (in terms of RE, see Fig. 9) of the reconstructions is better for lower levels than for higher levels, better for GPH than for wind (because one wind component often is much weaker than the other we plot only the higher regional average of the two components) and temperature, better in the boreal winter period compared to summer, and better in the northern extratropics compared to all other regions (see also BL2004, Griesser et al. 2010). There are interesting exceptions, however. The skill is higher in some cases for 700 than for 850 hPa, e.g. for temperature above Central Europe and the eastern USA in the 1950s, wind above tropical West Africa in the 1950s, or GPH above the eastern USA before 1945. The reason could be, e.g.,

inconsistencies between observations and calibration data (ERA-40) in the boundary layer or the monsoon layer, such as those found for the NNR by Stickler and Brönnimann (submitted manuscript). In the subtropical North Pacific region the skill for temperature and GPH is even better in the upper troposphere than in the lower troposphere. Again, a possible reason could be systematic differences between calibration data and observations, e.g., with respect to high subsidence inversions.

In Central Europe we see a clear increase of the skill during the Second World War and a sharp decrease in the post-war years, reflecting the amount of available upper-air data. In eastern North America, a substantial increase in the radiosonde network around 1945 produces much better skill in the stratosphere. All in all, the skill is high over the densely sampled regions of Europe and North America, where all fields and variables (except the lower stratosphere in summer) exhibit very high RE values. Reconstructions in tropical Africa and (for wind) the Russian Arctic perform badly and will likely not lead to useful interpretations.



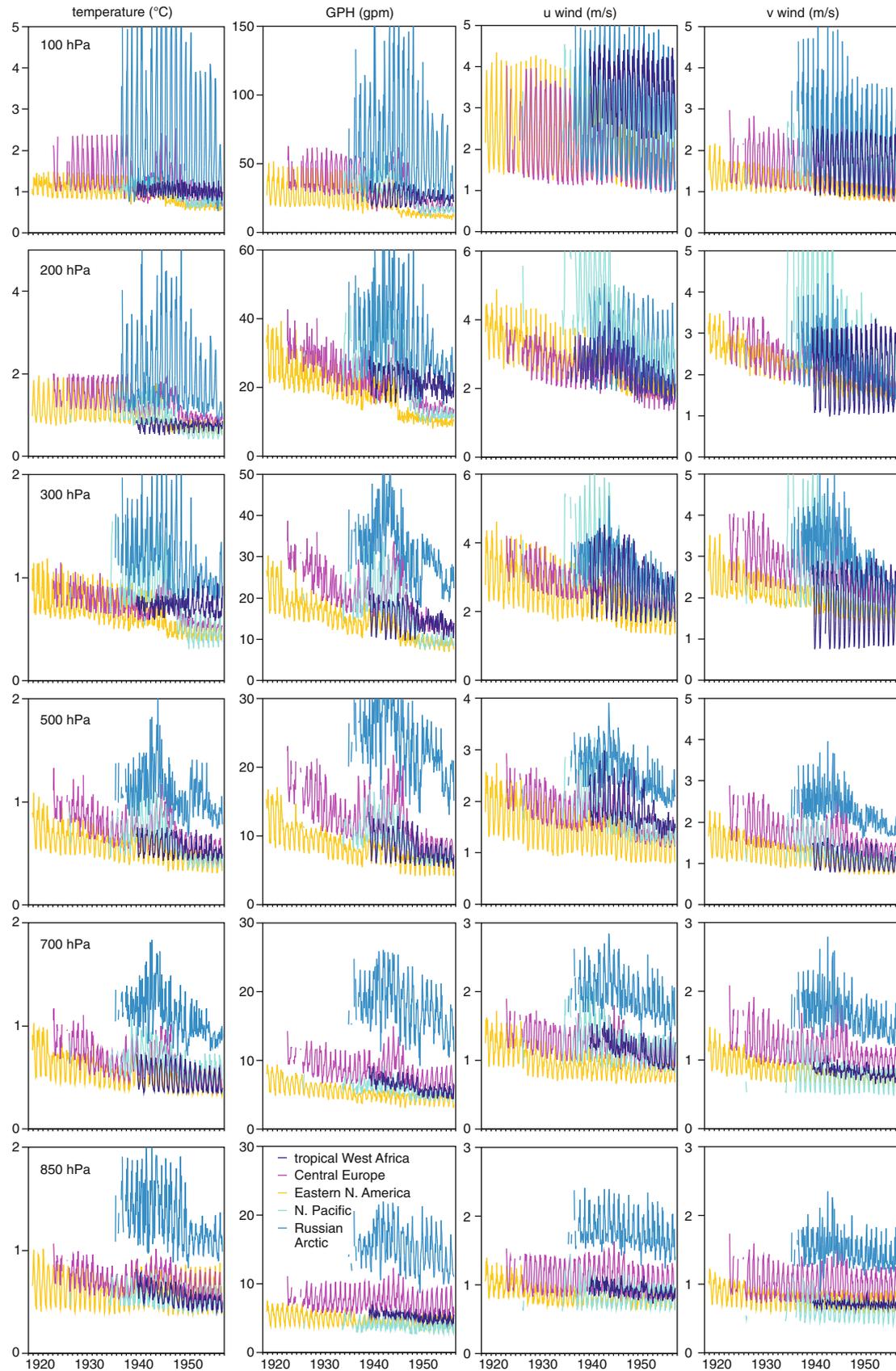
**Fig. 7** RE of the reconstructed temperature at 700 hPa in January 1920, January 1930, January 1940, and January 1950, respectively

A comparison of the new reconstruction with existing reconstructions and the 20CR is given in Fig. 10 in the form of correlation maps that are based on monthly anomalies for the example of 300 hPa GPH. We use three periods namely 1930–1938 where we compare REC2 with the reconstructions over Europe by Schmutz et al. (2001) as well as with 20CR, 1939–1947 where we compare REC2 with BL2004 and 20CR, and 1948–1957 where we compare REC2 with NNR and 20CR. Similar as to what has been found in the previous paragraphs and in 20CR (Compo et al. 2010), correlations are very high at northern midlatitudes, but decrease towards the tropics and also towards the polar regions. Correlations exceed 0.98 over large areas of the midlatitudes and sometimes even lie above 0.995. Even over the tropics, correlations do not drop below 0.5 when compared with NNR, which is promising for future work.

Interestingly, correlations in the first period are slightly higher for 20C than for the Schmutz et al. (2001) reconstruction, whereas in the second and third period, they are somewhat lower for 20CR compared with BL2004 and

NNR, respectively. However, the differences are very small. In all, this figure confirms that the results from the split-sample validation are qualitatively consistent with the results from a comparison with other studies. Quantitatively it shows that for GPH, an excellent quality can be expected over the midlatitudes.

Independent validation can also be performed with historical ozone data (see also Griesser et al. 2010). Because at midlatitudes the ozone column is strongly controlled by redistribution processes in the lower stratosphere and hence atmospheric dynamics, strong correlations with GPH and temperature in the upper troposphere and lower stratosphere are expected. Figure 11 shows for a number of locations the correlation profiles between historical total ozone data and reconstructed temperature (grey) and GPH (black) at all levels (solid) calculated on the basis of monthly anomalies from the ERA-40 period mean annual cycle. For comparison, the same profiles are also given for the correlation in a later period (9/1957–8/2002) using ERA-40 and ground-based or satellite total ozone data (dashed). For the two European stations Arosa and Oxford (with data from 1924 to 1957), we



**Fig. 8** Time series of averaged RMSE for five regions (denoted with *squares* in Fig. 4) for all levels and variables

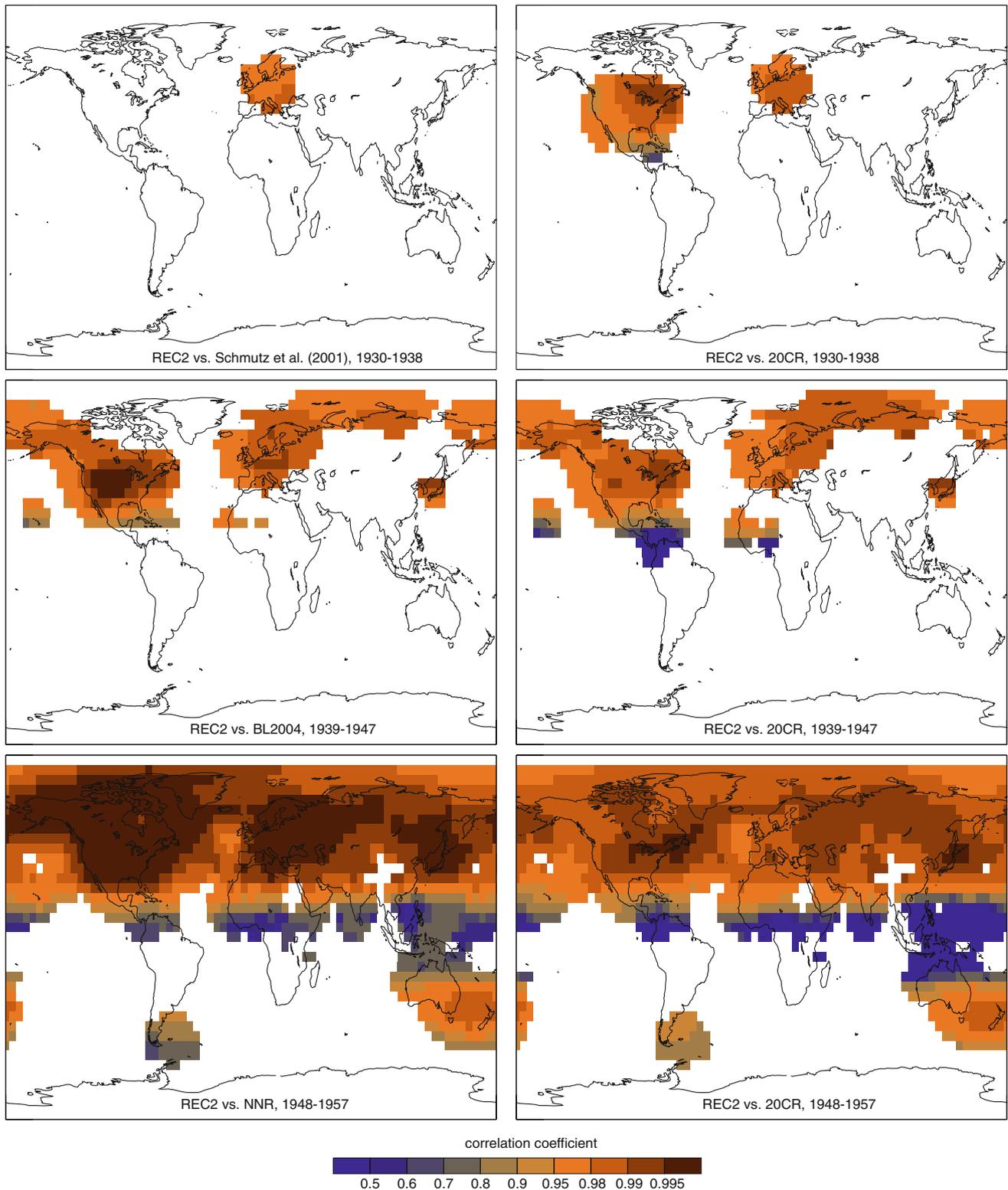


**Fig. 9** Time series of averaged RE for five regions (denoted with *squares* in Fig. 4) for all levels and variables. For wind we plot the maximum of the area-averaged  $u$  and  $v$  winds. Note that the y-axis is transformed using the function  $1 - \ln(1 - RE)$

clearly find that the correlation structure is well reproduced even in the stratosphere, albeit the absolute values of the correlations are lower than in the later period. The imperfect reconstructions likely also contribute to the worse correlations in the historical period, but so do the imperfect total ozone data. A very good comparison is found for New York in the 1940s. In the cases of Tromsø (1935–1957) and Zi-Ka-Wei (Shanghai, 1932–1942), the number of historical data pairs is lower and the shape of the profile less well specified than in the later period. Nonetheless, the correlations are strong (in many cases even stronger than in the later period).

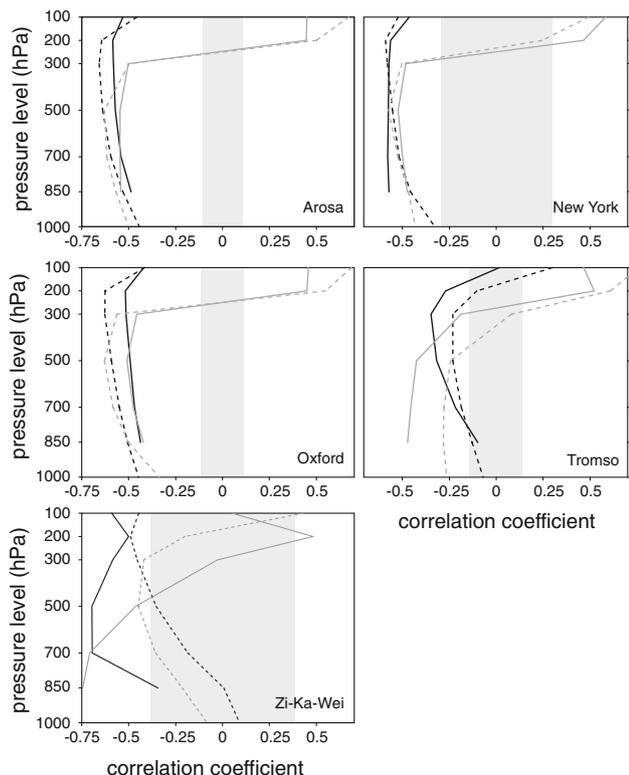
This is interesting in view of the fact that these two stations represent areas with a rather low density of upper-air information.

In all, based on the split-sample validations, the comparisons with other data products, and the comparison with fully independent total ozone data, REC2 is expected to meet the requirements of many applications, but we also identified instances where this is not the case. The error measures from the split-sample validations are made available together with the data and give the user the possibility to judge the quality of the data.



**Fig. 10** Correlation of monthly anomalies (from the mean seasonal cycle over the indicated time period) of 300 hPa GPH between REC2 and other data sets. *Top* Schmutz et al. (2001) and 20CR for the period 1930–1938, *middle* BL2004 and 20CR for the period 1939–1947, *bottom* NNR and 20CR for the period 1948–1957.

Correlations were based on anomalies from the mean seasonal cycle over the respective period (no missing values are allowed for plotting a correlation coefficient) after interpolating the data to a common  $5^\circ \times 5^\circ$  grid

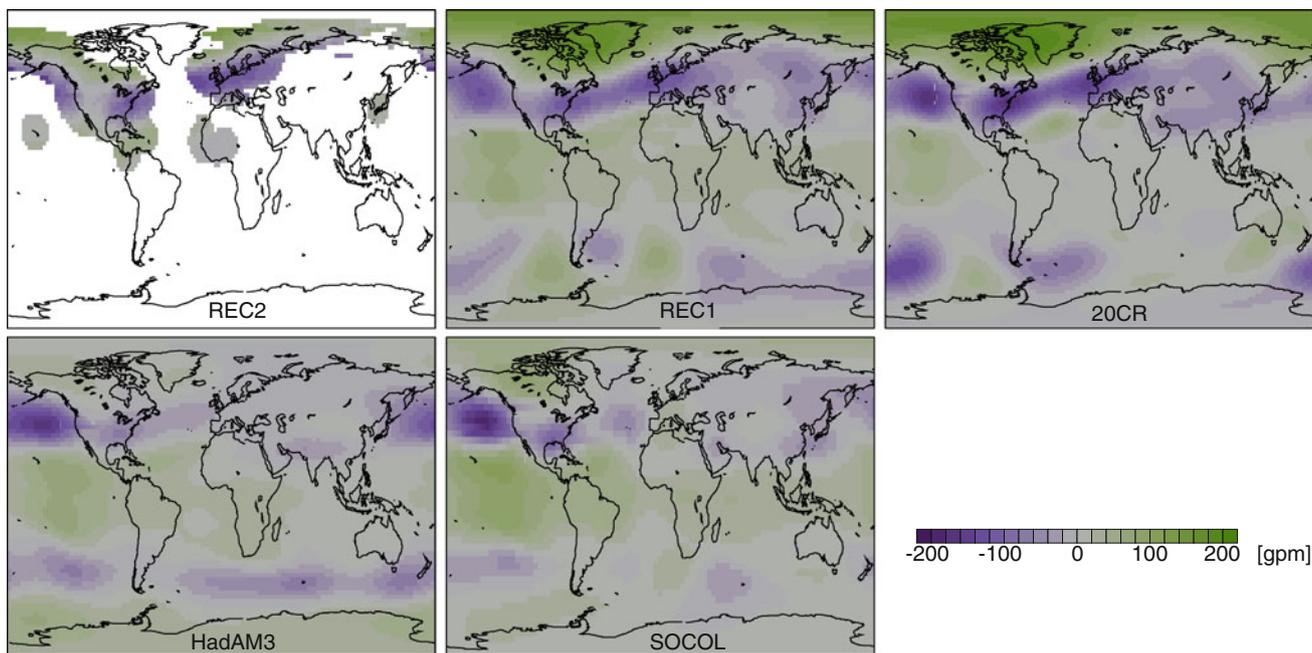


**Fig. 11** Profiles of the correlation between historical total ozone data and temperature (*grey*) and GPH (*black*) in the reconstruction period (*solid*) as well as profiles of the correlation between TOMS total ozone and ERA-40 temperature and GPH (*dashed*). *Grey shading* denotes insignificant correlation ( $\alpha = 0.05$ )

### 4 Examples

In this section we would like to show two cases, for which the data could be used. We chose two relatively well studied cases to be able to pinpoint the advantages and disadvantages of the new data sets: namely the 1940–1942 global climate anomalies (which were related to an El Niño event) and the Midwest US Dust Bowl droughts in 1931–1939.

The 1940–1942 global climate anomaly was arguably one of the strongest large-scale climate signals on inter-annual time scales in the twentieth century (see Brönnimann et al. 2004 and Fischer et al. 2008a, for more details). A particularly large signature is found in the northern extratropics (see Brönnimann et al. 2004). We therefore consider it necessary that data sets agree on this event. We focus on the 200 hPa level, which allows addressing both the upper-tropospheric planetary wave structure in the extratropics and the polar vortex in the stratosphere. Figure 12 shows 200 hPa GPH for the El Niño winters (JFM 1940, 1941, and 1942) minus the two following La Niña winters (JFM 1943, 1944) in five different data sets: three observation based data sets and two sets of model simulations. The general pattern is very clear: GPH was increased over the eastern tropical Pacific and over the polar region, while a band with decreased GPH, with a clear wave structure, is found at midlatitudes. Note that 20CR and REC1 are almost indistinguishable in the



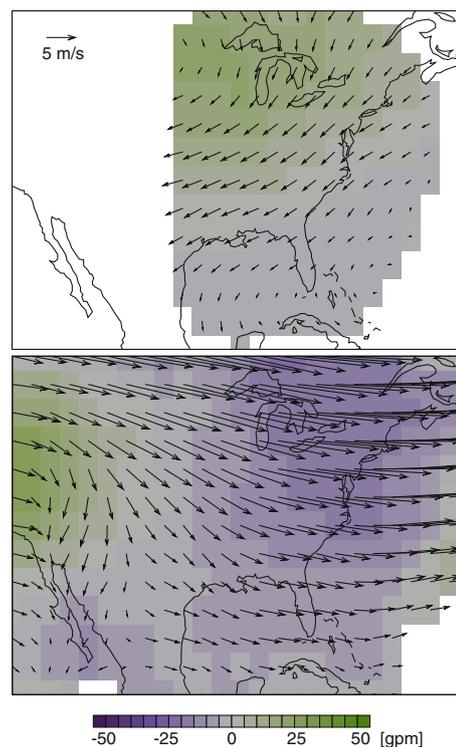
**Fig. 12** Difference in 200 hPa GPH between January–March 1940–1942 and January–March 1943–1944 in various data sets. HadAM3 and SOCOL are 6- and 9-member ensemble mean values, respectively

northern extratropics. Slight differences arise in the tropics and very large difference in the southern hemisphere. REC2 fits extremely well with both REC1 and 20CR. It captures the gradient in the anomaly field very well and thus confirms that the main pattern in the 200 hPa GPH in REC1 field is not simply the effect of the large-scale PC analysis used but can be well constrained locally. On the other hand, REC2 does not provide the full spatial coverage (no missing values were allowed for this plot, i.e., all 15 months need valid data). Some differences are seen over Japan.

The agreement is surprising in view of the fact that the data sets are partly or fully independent. 20CR includes information from SLP as well as monthly SST data. 20CR and the models have SSTs and sea ice as a common input field, while 20CR and reconstructions have SLP as a common input. Reconstructions and model are completely independent. Only REC1 and REC2 have upper-air data. The comparison with the model simulations shows that SSTs alone are sufficient to determine 200 hPa GPH over the tropical Pacific as well as over the Pacific-North American sector. The downstream signal over the North Atlantic-European sector is only depicted in a qualitative sense and the weak polar vortex only in SOCOL. Obviously atmospheric data is needed to reproduce these features.

The second example is the Dust Bowl droughts in the US Midwest in 1931–1939. Figure 13 shows 850 hPa winds and 500 hPa GPH for the summer (April–August) season in 1931–1939 compared to two reference period. In a previous paper (Brönnimann et al. 2009a) we analysed the Great Plains Low Level Jet directly in the historical wind data from the CHUAN data set (Stickler et al. 2010). We were able to address subtle features in the vertical structure and spatial extent, thereby accounting for the very large diurnal cycle by constraining the analysis to a short diurnal time window. However, working with the data directly forced us to compare the Dust Bowl years (1931–1939) with the subsequent wet years (1941–1944) only, rather than with a longer climatology such as 1921–1950. This was due to the heterogeneity of the data (lack of continuous series) rather than the amount of data.

With REC2 we cannot address the subtle features in the low-level jet. The vertical resolution is low, the lowermost levels may be not well represented (because ERA-40 and the local observations might systematically disagree), and no diurnal cycle is depicted (which may be crucial as the low-level jet is a nocturnal phenomenon). However, REC2 is able to produce a more homogeneous data set despite the heterogeneous wind data and thus allows using a longer reference period (Fig. 13). If we use the very wet period 1941–1944 as a reference we find results that are consistent with those obtained directly from the wind data, indicating



**Fig. 13** Difference in 850 hPa winds and 500 hPa GPH between April–August 1931–1939 and (top) 1921–1950 and (bottom) 1941–1944 in REC2

northwesterly to northerly anomalies in the lower troposphere during the Dust Bowl (Fig. 3, Brönnimann et al. 2009a). If, in turn, the Dust Bowl is expressed as anomalies from 1921 to 1950, the high pressure anomaly is shifted towards the east, and the wind over the Great Plains shows northerly to northeasterly anomalies. This is in agreement with the results of the 500 hPa reconstructed GPH from REC1 (reference period 1921–1950) and with an analysis of dry compared to wet years from ERA-40 (Figs. 3 and 4, Brönnimann et al. 2009a). The weakening of the Low Level Jet remains, though. This example shows advantages and disadvantages of the new data set as compared to analyzing the observations directly. REC2 draws from the full information available, not just the continuous records. Thus, it gives a more complete picture and allows better comparisons in time at the price of subtle details that can be addressed only in the observations directly.

These two examples demonstrate the strengths and weaknesses of the new data set in comparison with other data sets. One important conclusion is that no data set is preferable for all situations. We recommend to use no one, but all data sets for studying the first half of the twentieth century in order to account for the different shortcomings in the individual data sets.

## 5 Conclusion

In this paper we describe a new global gridded 3-D monthly atmospheric data set back to 1918, termed REC2. The data set is based on historical upper-air data as much as possible and uses statistical reconstruction techniques in a more modest way than previous approaches. An overview of the main features and the difference to the previous reconstruction REC1 is given in Table 1. By defining a “cone of influence” around each grid column and excluding information from outside that cone to affect the reconstructions, the stationarity assumption is alleviated as much as possible and becomes a local stationarity, that is more physical and easier to defend. The main disadvantage is the incomplete spatial coverage.

The data set has been evaluated using cross-validation methods, comparison with independent data and comparison with other reconstructions and climate models. Results show that the reconstruction skill varies with region, season, level, variable, and time. At northern midlatitudes, the reconstruction skill is generally good up to the upper troposphere, regionally also in the lower stratosphere. The skill is worse in some tropical regions such as Africa. The data set is suitable for studying interannual variability of the large-scale circulation and its effect on climate (e.g., droughts, ENSO influence), but it has not been validated for trend analysis.

The data set is made available through our platform <http://www.historicalupperair.org> together with error measures. We recommend that REC2 is used together with all other available data sets of this period (20CR, REC1, CHUAN, BL2004) and, most importantly, together with information on the errors and the generation of each data set. More work needs to be done to extend global three-dimensional data sets back to the first half of the twentieth century in order to obtain an overarching view of climate variability during the first half of the twentieth century.

**Acknowledgments** This work was supported by the Swiss National Science Foundation, Project “Past climate variability from an upper-level perspective”. We also gratefully acknowledge the support of the Swiss National Centre of Competence in Research on Climate (NCCR-Climate), funded by the Swiss National Science Foundation. We wish to thank Gil Compo, Jeff Whitaker and Prashant Sardeshmukh (University of Colorado and NOAA) for providing 20CR data, ECMWF for providing the ERA-40 data and the Hadley Centre of the UK MetOffice for providing the HadCRUT3v and the Had-SLP2 data sets.

## Appendix: Addition of noise to the ERA-40 predictors

Gaps in the predictor data (i.e., the observations) during the calibration period were filled with ERA-40 data. In order to account for the differences between reanalysis and observations, we perturbed the interpolated reanalysis data (i.e., all filled gaps) by a stochastic error model with three

components: a constant bias, an error in each monthly profile, and a random error. Except for the last error, the errors within a profile are dependent. Therefore, we first defined error profiles which in a later step were scaled with random numbers. The error is  $0.75^{\circ}\text{C}$  for temperature,  $1\text{ m/s}$  for wind (constant with altitude), and 12.3, 14.8, 16.8, 19.2, 22.1, 25.9, 31.1, 35.1, and  $40.1\text{ gpm}$ , respectively, for GPH at 850, 700, 600, 500, 400, 300, 250, 200, and 100 hPa.

The introduced bias is a step inhomogeneity with random start date (sampled from an equal distribution) and random length (between 5 and 30 years) that affects each station independently. Note that if the ending date is still within the period, there will be two steps. The bias is thought to describe step inhomogeneities that may arise due to changes in instruments, for instance. The bias is vertically coherent, i.e., we sampled one number from a random normal distribution with a mean of 0 and a standard deviation of 1,  $N(0,1)$ , to scale the error profiles of both variables (i.e., either  $u$  and  $v$  or GPH and temperature; there is no record with all four variables). Large temperature errors (constant with altitude) are accompanied by large GPH errors (increasing in magnitude with height) of the same sign. This is a very typical error for radiosonde data (although other errors are possible). Large  $u$ -wind anomalies are accompanied by large  $v$ -wind anomalies of the same sign, which is also an error that can occur (e.g., due to unit errors or errors in the reduction process), although there is no such thing as a typical wind error profile.

The profile error also is vertically coherent (as described above), but applies to each individual monthly mean independent of the previous month. It was also generated by sampling from  $N(0,1)$ . Errors of that sort may arise from sampling, but also from instrumental errors. Finally, the random error is completely random, i.e., has no vertical, spatial, or temporal structure. This part of the error measures the random error of the instrument reading. It was also generated by sampling from  $N(0,1)$ , but in this case for each variable and level individually.

The contributions of the three sources of error to the total variance of the error were chosen as 25, 37.5, and 37.5%. Note that these fractions as well as the timing of biases (see above) are somewhat arbitrarily chosen. Real observational series may have more complex step inhomogeneities (or network-wide biases), they may also have trend inhomogeneities, and  $u$  and  $v$  wind errors may have different relations. Also, possible undercorrection or overcorrection of radiation errors could lead to errors that are dependent on the altitude, the time of day, and the month of the year. However, the assumptions required for modeling such errors are increasingly difficult to justify. In any case, the disturbances resulting from our process have

the desired magnitudes (see Brönnimann 2003; Grant et al. 2009a, b, and Stickler et al. 2010 for approximate errors in the historical upper-air data) and appear reasonable.

Note that these errors adjust for the differences between ERA-40 and upper-air observations and hence they also represent (partly unknown) errors and uncertainties in ERA-40 (although at locations where upper-air data was assimilated into ERA-40, the two are normally very close).

## References

- Allan R, Ansell T (2006) A new globally complete monthly historical gridded mean sea level pressure dataset (HadSLP2): 1850–2004. *J Clim* 19:5816–5842
- Bengtsson L, Hagemann S, Hodges KI (2004) Can climate trends be calculated from reanalysis data? *J Geophys Res* 109:D11111. doi:10.1029/2004JD004536
- Bromwich D, Wang S (2005) Evaluation of the NCEP-NCAR and ECMWF 15- and 40-yr reanalyses using rawinsonde data from two independent Arctic field experiments. *Mon Wea Rev* 133:3562–3578
- Brönnimann S (2003) A historical upper-air data set for the 1939–1944 period. *Int J Climatol* 23:769–791
- Brönnimann S, Luterbacher J (2004) Reconstructing Northern Hemisphere upper-level fields during World War II. *Clim Dyn* 22:499–510
- Brönnimann S, Luterbacher J, Schmutz C, Wanner H, Staehelin J (2000) Variability of total ozone at Arosa, Switzerland, since 1931 related to atmospheric circulation indices. *Geophys Res Lett* 27:2213–2216
- Brönnimann S, Cain JC, Staehelin J, Farmer SFG (2003) Total ozone observations prior to the IGY. II. Data and quality. *Q J Roy Meteorol Soc* 129:2819–2843
- Brönnimann S, Luterbacher J, Staehelin J, Svendby TM, Hansen G, Svenøe T (2004) Extreme climate of the global troposphere and stratosphere in 1940–42 related to El Niño. *Nature* 431:971–974
- Brönnimann S, Stickler A, Griesser T, Ewen T, Grant AN, Fischer AM, Schraner M, Peter T, Rozanov E, Ross T (2009a) Exceptional atmospheric circulation during the “Dust Bowl”. *Geophys Res Lett* 36:L08802. doi:10.1029/2009GL037612
- Brönnimann S, Stickler A, Griesser T, Fischer AM, Grant A, Ewen T, Zhou T, Schraner M, Rozanov E, Peter T (2009b) Variability of large-scale atmospheric circulation indices for the Northern Hemisphere during the past 100 years. *Meteorol Z* 18:379–396
- Compo GP et al (2010) The twentieth century reanalysis project. *Q J R Meteorol Soc* (submitted)
- Cook ER, Briffa KR, Jones PD (1994) Spatial regression methods in dendroclimatology—a review and comparison of two techniques. *Int J Climatol* 14:379–401
- Cook BI, Seager R, Miller RL (2010) Atmospheric circulation anomalies during two persistent North American droughts: 1932–1939 and 1948–1957. *Clim Dyn* (online first). doi:10.1007/s00382-010-0807-1
- Durre I, Vose RS, Wuertz DB (2006) Overview of the Integrated Global Radiosonde Archive. *J Clim* 19:53–68
- Emanuel K (2010) Tropical cyclone activity downscaled from NOAA-CIRES reanalysis, 1908–1958. *J Adv Model Earth Syst* 2:1. doi:10.3894/JAMES.2010.2.1
- Ewen T, Grant A, Brönnimann S (2008a) A monthly upper-air data set for North America back to 1922 from the Monthly Weather Review. *Mon Wea Rev* 136:1792–1805
- Ewen T, Brönnimann S, Annis JL (2008b) An extended Pacific North American index from upper air historical data back to 1922. *J Clim* 21:1295–1308
- Fischer AM, Shindell DT, Winter B, Bourqui MS, Faluvegi G, Rozanov E, Schraner M, Brönnimann S (2008a) Stratospheric winter climate response to ENSO in three chemistry-climate models. *Geophys Res Lett* 35:L13819. doi:10.1029/2008GL034289
- Fischer AM et al (2008b) Interannual-to-decadal variability of the stratosphere during the 20th century: ensemble simulations with a Chemistry-Climate Model. *Atmos Chem Phys* 8:14371–14418
- Free M, Seidel DJ, Angell J, Lanzante JK, Durre I, Peterson TC (2005) Radiosonde atmospheric temperature products for assessing climate (RATPAC): a new data set of large-area anomaly time series. *J Geophys Res* 110:D22101. doi:10.1029/2005JD006169
- Giese BS, Compo GP, Slowey NC, Sardeshmukh PD, Carton JA, Ray S, Whitaker JS (2010) The 1918/1919 El Niño. *Bull Am Meteorol Soc* 91:177–183. doi:10.1175/2009BAMS2903
- Grant A, Brönnimann S, Haimberger L (2008) Recent Arctic warming vertical structure contested. *Nature* 455:E2–E3. doi:10.1038/nature07257
- Grant A, Brönnimann S, Ewen T, Nagurny A (2009a) A new look at radiosonde data prior to 1958. *J Clim* 22:3232–3247
- Grant AN, Brönnimann S, Ewen T, Griesser T, Stickler A (2009b) The early twentieth century warm period in the European Arctic. *Met Z* 18:425–432
- Griesser T, Brönnimann S, Grant A, Ewen T, Stickler A, Comeaux J (2010) Reconstruction of global monthly upper-level temperature and geopotential height fields back to 1880. *J Clim* (early online release). doi:10.1175/2010JCLI3056.1
- Haimberger L (2007) Homogenization of radiosonde temperature time series using innovation statistics. *J Clim* 20:1377–1403
- Haimberger L, Tavolato C, Sperka S (2008) Towards the elimination of warm bias in historic radiosonde records—some new results from a comprehensive intercomparison of upper air data. *J Clim* 21:4587–4606
- Hansen G, Svenøe T (2005) Multilinear regression analysis of the 65-year Tromsø total ozone series. *J Geophys Res* 110:D10103. doi:10.1029/2004JD005387
- Hansen J, Ruedy R, Glascoe J, Sato M (1999) GISS analysis of surface temperature change. *J Geophys Res* 104:30997–31022
- Harnik N, Chang EKM (2003) Storm track variations as seen in radiosonde observations and reanalysis data. *J Clim* 16:480–495
- Jones PD, Wigley TML, Briffa KR (1987) Monthly mean pressure reconstructions for Europe (back to 1780) and North America (to 1858). U.S. Dept. of Energy Carbon Dioxide Research Division, Technical Report TRO37, 99 pp
- Kistler R et al (2001) The NCEP-NCAR 50-year reanalysis: monthly means CD-ROM and documentation. *Bull Am Meteorol Soc* 82:247–267
- Luterbacher J, Xoplaki E, Dietrich D, Rickli R, Jacobeit J, Beck C, Gyalistras D, Schmutz C, Wanner H (2002) Reconstruction of sea level pressure fields over the Eastern North Atlantic and Europe back to 1500. *Clim Dyn* 18:545–561
- Luterbacher J, Dietrich D, Xoplaki E, Grosjean M, Wanner H (2004) European seasonal and annual temperature variability, trends, and extremes since 1500. *Science* 303:1499–1503. doi:10.1126/science.1093877
- Santer BD et al (2004) Identification of anthropogenic climate change using a second-generation reanalysis. *J Geophys Res* 109:D21104. doi:10.1029/2004JD005075
- Scaife AA et al (2009) The CLIVAR C20C Project: selected 20th century climate events. *Clim Dyn* 33:603–614
- Schmutz C, Gyalistras D, Luterbacher J, Wanner H (2001) Reconstruction of monthly 700, 500 and 300 hPa GPH fields in the

- European and Eastern North Atlantic region for the period 1901–1947. *Clim Res* 18:181–193
- Sherwood SC, Meyer CL, Allen RJ (2008) Robust tropospheric warming revealed by iteratively homogenized radiosonde data. *J Clim* 21:5336–5350
- Simmons AJ, Jones PD, da Costa Bechtold V, Beljaars ACM, Kallberg PW, Saarinen S, Uppala SM, Viterbo P, Wedi N (2004) Comparison of trends and low-frequency variability in CRU, ERA-40 and NCEP/NCAR analyses of surface air temperature. *J Geophys Res* 109:D24115. doi:[10.1029/2004JD005306](https://doi.org/10.1029/2004JD005306)
- Staehelin J, Renaud A, Bader J, McPeters R, Viatte P, Hoegger B, Buignion V, Giroud M, Schill H (1998) Total ozone series at Arosa (Switzerland): homogenisation and data comparison. *J Geophys Res* 103:5827–5841
- Stickler A et al (2010) The comprehensive historical upper air network (CHUAN). *Bull Am Meteorol Soc* 91:741–751. doi:[10.1175/2009BAMS2852.1](https://doi.org/10.1175/2009BAMS2852.1)
- Thorne PW, Parker DE, Tett SFB, Jones PD, McCarthy M, Coleman H, Brohan P (2005) Revisiting radiosonde upper-air temperatures from 1958 to 2002. *J Geophys Res* 110:D18105. doi:[10.1029/2004JD005753](https://doi.org/10.1029/2004JD005753)
- Trenberth KE, Stepaniak DP, Hurrell JW (2001) Quality of reanalyses in the tropics. *J Clim* 14:1499–1510
- Uppala SM et al (2005) The ERA-40 re-analysis. *Q J R Meteorol Soc* 131:2961–3012
- Vogler C, Brönnimann S, Staehelin J, Griffin REM (2007) The Dobson total ozone series of Oxford: re-evaluation and applications. *J Geophys Res* 112:D20116. doi:[10.1029/2007JD008894](https://doi.org/10.1029/2007JD008894)
- Wood KR, Overland JE (2010) Early 20th century Arctic warming in retrospect. *Int J Climatol* 30:1269–1279. doi:[10.1002/joc.1973](https://doi.org/10.1002/joc.1973)