A New Method for the Comparison of Trend Data with an Application to Water Vapor

SEBASTIAN MIERUCH, STEFAN NOEL, MAXIMILIAN REUTER, HEINRICH BOVENSCHMANN, AND JOHN P. BURROWS
Institute of Environmental Physics (IUP), University Of Bremen, Bremen, Germany

MARC SCHRODER
Satellite Application Facility on Climate Monitoring (CMSAF), German Weather Service, Offenbach, Germany

JORG SCHULZ
European Organisation for the Exploitation of Meteorological Satellites, Darmstadt, Germany

(Manuscript received 10 February 2010, in final form 12 January 2011)

ABSTRACT

Global total column water vapor trends have been derived from both the Global Ozone Monitoring Experiment (GOME) and the Scanning Imaging Absorption Spectrometer for Atmospheric Chartography (SCIAMACHY) satellite data and from globally distributed radiosonde measurements, archived and quality controlled by the Deutscher Wetterdienst (DWD).

The control of atmospheric water vapor amount by the hydrological cycle plays an important role in determining surface temperature and its response to the increase in man-made greenhouse effect. As a result of its strong infrared absorption, water vapor is the most important naturally occurring greenhouse gas. Without water vapor, the earth surface temperature would be about 20 K lower, making the evolution of life, as we know it, impossible. The monitoring of water vapor and its evolution in time is therefore of utmost importance for our understanding of global climate change. Comparisons of trends derived from independent water vapor measurements from satellite and radiosondes facilitate the assessment of the significance of the observed changes in water vapor.

In this manuscript, the authors have compared observed water vapor change and trends, derived from independent instruments, and assessed the statistical significance of their differences. This study deals with an example of the Behrens–Fisher problem, namely, the comparison of samples with different means and different standard deviations, applied to trends from time series.

Initially the Behrens–Fisher problem for the derivation of the consolidated change and trends is solved using standard (frequentist) hypothesis testing by performing the Welch test. Second, a Bayesian model selection is applied to solve the Behrens–Fisher problem by integrating the posterior probabilities numerically by using the algorithm Differential Evolution Markov Chain (DEMC). Additionally, an analytical approximative solution of the Bayesian posterior probabilities is derived by means of a quadratic Taylor series expansion applied in a computationally efficient manner to large datasets. The two statistical methods used in the study yield similar results for the comparison of the water vapor changes and trends from the different measurements, yielding a consolidated and consistent behavior.

1. Introduction

The estimation of robust parameters that describe change from time series of data and their statistical significance is one important aspect of modern climate research. Sophisticated and appropriate statistical methods are needed for this task. Data for climate studies is collected within large international projects, designed for time spans of several decades—for example, the Global Climate Observing System (GCOS). Such data includes both in situ measurements, such as radiosondes, and remote sensing measurements from spaceborne instruments. This effort...
has been generating long-term time series of diverse quantities (Schulz et al. 2009). The most famous climate time series ever recorded is probably the CO₂ time series quantities (Schulz et al. 2009). The most famous climate has been generating long-term time series of diverse water vapor because of their independence. However, the comparison and the assessment of significance of trends in Wetterdienst (DWD). These data are well suited for Monitoring (CMSAF), which is hosted by the Deutscher 2002. In addition, the water vapor trends have been used (Bovensmann et al. 1999)—on board tetrometer for Atmospheric Chartography (SCIAMACHY) since 1995—and the Scanning Imaging Absorption Spectrometer (GOME) (Burrows et al. 1999)—onboard 1998). To test the robustness of the estimated errors in the derived trends, and to quantitatively establish the representativeness of the trends, it is not sufficient to calculate only the trends and their errors, but it is also necessary to compare trends from independent instruments having overlapping measurements.

In this manuscript, we compare trends of total water vapor columns obtained from different sources. One dataset is from the Global Ozone Monitoring Experiment (GOME) (Burrows et al. 1999)—onboard ERS-2 since 1995—and the Scanning Imaging Absorption Spectrometer for Atmospheric Chartography (SCIAMACHY) (Bovensmann et al. 1999)—on board ENVISAT since 2002. In addition, the water vapor trends have been used from the globally distributed radiosonde network stations archived by the Satellite Application Facility on Climate Monitoring (CMSAF), which is hosted by the Deutscher Wetterdienst (DWD). These data are well suited for comparison and the assessment of significance of trends in water vapor because of their independence. However, the datasets have different sampling times, spatial resolution, etc. So, the issue arises how to best compare these data and the trends derived from them.

The appropriate comparison of data or quantities such as means and standard deviations from different measurement techniques is a common problem in a variety of scientific disciplines. One possible solution for this issue has a long tradition, going back at least to 1929, when Behrens (Behrens 1929) proposed a solution to the problem of the accurate determination of the difference of means in data from different sources, assuming that the standard deviations of the different datasets are not equal and unknown. Fisher (1937) found a solution to the problem using fiducial inference. The Behrens–Fisher problem has also been solved by Jeffreys (1939) and recently, for example, by Brethorst (1993) and Moreno et al. (1999), who have used Bayesian probability analysis. Currently, examples of the Behrens–Fisher problem are mostly solved by using the frequentist statistics approach—namely, the Welch test (Welch 1947)—which is an adaption of a Student's t test. Unfortunately, the Behrens–Fisher problem is often incorrectly simplified in such a way that a difference of given quantities is stated as significant if the respective error bars do not overlap, as discussed by Lanzante (2005).

In this study the Behrens–Fisher problem associated with the derivation of trends in water vapor from different time series having different, and possibly unknown, standard deviations has been investigated. The objective of our analysis is twofold. First, methods for trend comparisons in the sense of hypothesis testing under the frequentist and Bayesian framework have been developed. Second, the methods have been applied to water vapor data in a demonstration study, which shows that valuable meteorological information is gained successfully by applying the methods developed in this manuscript.

The manuscript is structured as follows. In section 2, we introduce the databases used, that is, the water vapor measurements from satellites and ground stations. In section 3, we briefly describe the two relevant schools of statistics—"the frequentist" (standard) approach and the Bayesian concepts—and explain their underlying philosophy. The statistical methods used are discussed in section 4, including the trend calculation, the Welch test for trends, the Bayesian model selection applied to trends, a sensitivity study regarding prior information, and the derivation of an approximation of the Bayesian method. The three approaches are applied to the GOME–SCIAMACHY and radiosonde water vapor time series and compared in section 5. Finally, the conclusions are given in section 6.

2. The water vapor measurements

a. GOME–SCIAMACHY

The global total water vapor column amounts used in the present study have been retrieved by the Air Mass Corrected Differential Optical Absorption Spectroscopy approach (AMC–DOAS) (Noël et al. 2004) from observation of the upwelling radiation at the top of the atmosphere in the visible range measured by the GOME instrument onboard the orbiting platform ERS-2, launched in April 1995, and the SCIAMACHY spectrometer onboard ENVISAT, launched in March 2002. These satellites fly in sun-synchronous orbits in descending node at a height of about 785 km. GOME crosses the equator at 1030 local time and has a spatial resolution of typically 40 km × 320 km: global coverage being achieved after
about three days. SCIAMACHY crosses the equator at 1000 local time, its spatial resolution being somewhat better, typically at 30 km x 60 km, whereas global coverage is achieved within six days. Since the instruments are dependent on sun light, no information on water vapor is retrieved during night, which results in no data being obtained during the polar night. Scenes that have too large cloud cover and regions with extreme elevation (e.g., Himalayas) are excluded from the analysis (Noël et al. 2004). The GOME–SCIAMACHY dataset is gridded on a spatial 0.5° x 0.5° lattice and then accumulated to monthly means. This yields a cloud-cleared monthly mean water vapor climatology between 1000 and 1030 local time.

b. Radiosondes

Radiosondes are devices carried on small balloons to heights around 35 km. Typically the parameters pressure, altitude, geographical position, temperature, relative humidity, and wind speed are recorded and sent to a ground station. Comprehensive information on different sensors, designs, calibration, etc. can be found, for example, in Elliott and Gaffen (1991), Garand et al. (1992), Lally (1985), and Vaisala (1989).

Under the framework of the World Meteorological Organization a network of globally distributed radiosonde stations perform regular observations, which have been quality-controlled and archived at the DWD. In this study, 187 high quality time series (from more than 900 stations in total) are considered. The measured water vapor profiles have been integrated up to 100 hPa to yield the total column amounts. Typically the radiosondes are launched daily at 0600, 1200, 1800, and 0000 local time. These data have been averaged to derive monthly means, which are comparable with those determined from the satellite data. The time span is selected to match the satellite time range and the comparison is performed for data from January 1996 to December 2007.

There are several differences between the two datasets. One of the most important differences is horizontal spatial resolution, which is relatively low, that is, poor (0.5° x 0.5°) for the satellite data and high (point measurements) for the radiosonde observations. Further differences in the water vapor data are expected because of the AMC–DOAS data being cloud cleared, whereas the radiosonde data contain all-sky measurements. The difference between clear-sky and all-sky relative humidity has been explored, for example, by Kahn et al. (2009) for the upper troposphere, where a large impact has been observed. Since we are dealing with the total column water vapor, these effects in the upper troposphere have only a minor contribution. However, Sohn and Bennartz (2008) have investigated the clear-sky bias for total column water vapor and found a bias of about 0.2 g cm⁻² for zonal means between clear-sky and all-sky data. Additional biases can be expected due to diurnal variations of water vapor, as the satellite measurements in principle present a morning climatology (equator crossing at about 1000 local time), whereas the radiosondes sample the complete day. To account for such biases we use individual offsets for each dataset in the regression procedure in section 4a.

In spite of the differences in sampling there are important similarities in both datasets, for example, the temporal resolution being monthly. The averaging to monthly means typically smears out small-scale fluctuations and the individual offsets account for biases, which makes a comparison possible between the two datasets.

3. The two schools of statistics

The frequentist statistics approach was primarily developed by, for example, Fisher, Neyman, and Pearson at the beginning of the twentieth century. The underlying philosophy of the frequentist statistics is the interpretation of an event probability as the limit of its relative frequency for a large number of trials. A major component of how statistics is used in environmental science is the hypothesis testing, which is used under the framework of induction to make decisions using experimental data. The basic concept of hypothesis testing is to set up a null hypothesis $H_0$, which is assumed to be true and an alternative hypothesis $H_1$, which is the complementary event of $H_0$. Then the probability of exceeding a value of a test statistic (according to $H_0$) is inferred. The null hypothesis is typically rejected if the observed probability is below a significance level of, for example, $\alpha = 0.05$. Such a case would confirm the alternative hypothesis.

Bayesian statistics has been developed by Bayes (1763) and de Laplace (1812). It is much older than the frequentist approach but was then largely forgotten until Jeffreys (1939) rediscovered the ideas of Bayes and de Laplace. The Bayesian concepts have undergone a renaissance in the late twentieth century, in part as a result of the increase of computational power. Influence on the Bayesian development in recent times has been contributed by, for example, Jaynes and Bretthorst (2003). An advantage of the Bayesian formalism is that it is based completely on probability theory (Jaynes and Bretthorst 2003), whereas the frequentist statistics represents rather a compilation of a large amount of tests and methods. Hypothesis testing can also be accomplished within the Bayesian framework. However,
Bayesian hypothesis testing is better described as a model selection procedure, that is, inferring which model or hypothesis has the higher probability to explain certain data or phenomena.

The major differences between frequentist and Bayesian statistics are as follows.

- **Philosophical difference:** The deep philosophical difference is that the parameters are fixed (but unknown) in the former and have some randomness in the form of a prior (or degree of belief) distribution in the latter. Data is used by Bayesians to update the prior knowledge in the form of a prior distribution, resulting in a posterior distribution that expresses the relative evidence of the parameters values given the data and the prior knowledge. In contrast, frequentists calculate statistics from the data to estimate the parameters and calculate the distribution of these statistics, given in (hypothetical) other datasets generated under the same model and assumed fixed parameter values. This can be elucidated by conditional probabilities. A conditional probability is the probability of an event \( X \) given the occurrence of another event \( Y \) and is denoted as \( P(X|Y) \). In the frequentist approach \( X \) could be data or a statistic derived from the data and \( Y \) a hypothesis, for example, a model with a particular parameter value that is assumed to have generated the data. In hypothesis testing frequentists then calculate the exceedance probability \( P(X > X_0|Y) \), where \( X_0 \) is the (test) statistic calculated from the data at hand. Bayesians can give the reverse, \( P(Y|X) \), the probability of the hypothesis given the data. The frequentists probability requires the notion of other (hypothetical) datasets generated from the same model and parameters, whereas the Bayesian probability conditions on the particular data at hand.

- **Prior information:** Bayesian methods utilize prior information about the truth of a hypothesis or parameter range, which reflects the knowledge (or ignorance) before the data have been analyzed. This enables a new quality of statistical inference, for example, the investigation of the evidence for human-induced climate change, as performed by Lee et al. (2005). In frequentist statistics such prior information does not exist.

The fundament of Bayesian statistics is given by the Bayes theorem, which can be formulated as

\[
P(Y|X,I) = \frac{P(X|Y,I)P(Y|I)}{P(X|I)},
\]

where \( X \) and \( Y \) are propositions, and \( I \) denotes the relevant background information. The \( I \) is often neglected, but it has to be kept in mind that no absolute probabilities exist without certain background assumptions or information. The \( P(Y|X,I) \) is called the posterior probability; \( P(X|Y,I) \) is the likelihood; \( P(Y|I) \) is the prior probability; and \( P(X|I) \) has formerly been called the marginalization likelihood, for which Sivia and Skilling (2006) have introduced the term “evidence.”

4. Methods

a. Estimating trends from time series

As one quality criterion, the time series have to include at least two-thirds of all (144) data points considered for the comparison. This is because large data gaps are not representative. The disadvantage of the two-thirds criterion is that only 187 radiosonde time series fulfill this requirement, whereas the satellite data fulfill the criterion in all 908 cases.

Global GOME–SCIAMACHY water vapor trends have been calculated for the time span from 1996 to 2006 in Mieruch et al. (2008), where the methods have been adopted and slightly expanded from Weatherhead et al. (1998). We have extended the trend analysis, which now includes the time from 1996 to 2007.

In the following the water vapor trend estimation is shortly discussed. For a more detailed description we refer to Mieruch et al. (2008) and Weatherhead et al. (1998).

An individual GOME–SCIAMACHY time series (single grid point) can be described by the trend model shown in Eq. (2):

\[
Y_{1t} = \mu_1 C_{1t} + S_{1t} + \omega_1 X_{1t} + \delta U_t + N_{1t},
\]

where \( Y_{1t} \) contains the monthly mean water vapor data, \( \mu_1 \) is a constant to be estimated, and \( C_{1t} \) equals unity for all \( t \) and is needed when assessing autocorrelations; \( S_{1t} \) is the seasonal component, \( \omega_1 \) represents the trend, and \( X_{1t} \) contains the time. The subscript “1” is used for the satellite data, whereas we will use the subscript “2” for the radiosonde trend model.

To account for the change from the GOME to the SCIAMACHY data, that is, the change of spatial resolution and the time of measurement, we use a level shift (as suggested by Weatherhead et al. 1998) of fitted magnitude \( \delta \) at and after time \( t = T_0(1 < T_0 < T_1) \), where \( T_0 = 85 \) represents the intersection of GOME and SCIAMACHY data on January 2003. Here \( U_t \) describes a step function:

\[
U_t = \begin{cases} 0, & t < T_0 \\ 1, & t \geq T_0 \end{cases}
\]
The noise $N_{1t}$ is modeled as an autoregressive process of order one [AR(1)], that is, $N_{1t} = \phi_1 N_{1,t-1} + \epsilon_{1t}$ (Schlütgen and Streitberg 1997) to consider autocorrelations in the data, where $\epsilon_{1t}$ is an independent random variable with zero mean and variance $\sigma_1^2$. The magnitude of autocorrelation $\phi_1$ is restricted to $-1 < \phi_1 < 1$ and is estimated using the discrete autocorrelation function (Edelson and Krolik 1988), which can account for gaps in the data.

In the publications of Mieruch et al. (2008) and Weatherhead et al. (1998) the seasonal component is described as a Fourier series and subtracted from the measurements to yield deseasonalized data. For simplicity anomalies are calculated by subtracting the seasonal means from 1996 to 2007 from the monthly mean water vapor time series in this manuscript. In Mieruch (2010) the trends are shown to be invariant with respect to the choice of the two methods for deseasonalizing. Accordingly, we add the respective overall means to the data. The magnitude of the level shift $\delta$ and the trend $\omega_1$ are invariant under the calculation of anomalies. Thus, the deseasonalized GOME–SCIAMACHY water vapor is now modeled by

$$D_{1t} = \mu_1 C_{1t} + \omega_1 X_{1t} + \delta U_t + N_{1t},$$  \hspace{1cm} \textit{t} = 1, \ldots, T_1. \hspace{1cm} (4)

As shown by Mieruch et al. (2008) the autocorrelated noise $N_{1t}$ is transformed to white noise $\epsilon_{1t}$ using the AR(1). The autocorrelations are accommodated by the variables in Eq. (4) (now indicated with the asterisk superscript), thus we observe

$$D_{1t}^* = \mu_1 C_{1t}^* + \omega_1 X_{1t}^* + \delta U_t^* + \epsilon_{1t},$$ \hspace{1cm} \textit{t} = 1, \ldots, T_1. \hspace{1cm} (5)

Comparing the trend results derived here with those shown in Mieruch et al. (2008) results in small differences, indicating that the expansion of the data to 2007 influences the trends very slightly. The regression parameters $\mu_1$, $\omega_1$, and $\delta$ have been estimated by a least squares approach. Furthermore, the standard errors of the regression parameters $\sigma_{\mu_1}$, $\sigma_{\omega_1}$, and $\sigma_{\delta}$ and the standard deviation of the noise $\sigma_1$ have been determined within the regression procedure. Note that the standard error refers to the standard deviation divided by the sample size. Since we are interested in trends, only the error of the trend and the standard deviation of the noise are needed for the hypothesis testing. The trend model for the radiosonde measurements $D_{2t}^*$, where also autocorrelations have been considered, is

$$D_{2t}^* = \mu_2 C_{2t}^* + \omega_2 X_{2t}^* + \epsilon_{2t}, \hspace{1cm} \textit{t} = 1, \ldots, T_2, \hspace{1cm} (6)$$

where no level shift is used. In the same way as above, the regression parameters $\mu_2$, $\omega_2$, their errors $\sigma_{\mu_2}$, $\sigma_{\omega_2}$, and the standard deviation of the noise $\sigma_2$ have been estimated. For the approximation of the Bayesian method, shown in section 4c, the standard deviation of the noise from the pooled data with a single trend is needed, which can be obtained by applying the least squares regression to the pooled data $D_{pt}^* = [D_{1t}^*, D_{2t}^*]$:

$$D_{pt}^* = \mu_{p1} C_{p1t}^* + \mu_{p2} C_{p2t}^* + \omega_p X_{pt}^* + \delta_t U_{pt}^* + \epsilon_{pt},$$

\hspace{1cm} \textit{t} = 1, \ldots, T_1 + T_2, \hspace{1cm} (7)

with

$$C_{p1t}^* = \begin{cases} C_{1t}^*, & t \leq T_1 \\ 0, & t > T_1 \end{cases}, \hspace{1cm} C_{p2t}^* = \begin{cases} C_{2t}^*, & t \leq T_1 \\ 0, & t > T_1 \end{cases};$$

$$X_{pt}^* = \begin{cases} X_{1t}^*, & t \leq T_1 \\ X_{2t}^*, & t > T_1 \end{cases}, \hspace{1cm} U_{pt}^* = \begin{cases} U_{t}^*, & t \leq T_1 \\ 0, & t > T_1 \end{cases}. \hspace{1cm} (8)$$

To pool the data individual offsets are used in the regression model, being implemented as $\mu_{p1}$ and $\mu_{p2}$ in Eq. (7).

Solving the least squares regression, the standard deviation of the noise

$$\sigma_{p1} = \sqrt{\frac{1}{T_1 - 3} \sum_{t=1}^{T_1} \epsilon_{pt}^2},$$ \hspace{1cm} (9)

and

$$\sigma_{p2} = \sqrt{\frac{1}{T_2 - 2} \sum_{t=T_1+1}^{T_1+T_2} \epsilon_{pt}^2},$$ \hspace{1cm} (10)

under a single trend is determined. The $l_i$ are the respective lengths of the time series $D_{1t}^*$ and $D_{2t}^*$, which are reduced by the number of fitted parameters.

b. Welch test applied to trends

The Welch test (Welch 1947) is the most commonly used approach to solve the Behrens–Fisher problem, that is, the estimation of the probability of equal means with different unknown standard deviations from time series. It represents an unpaired $t$ test; thus we assume the independence of satellite and radiosonde measurements. This is a reasonable assumption regarding the differences between the datasets, which have been introduced in section 2. In the following, the Welch test is applied to trends from time series. The null hypothesis $H_0$: $d = \omega_1 - \omega_2 = 0$ postulates that the difference of the two trends is equal to zero, whereas the alternative hypothesis is $H_1$: $d = \omega_1 - \omega_2 \neq 0$. The standard error of the difference $d$ is observed as
\[ \sigma_d = \sqrt{\sigma_{\omega_1}^2 + \sigma_{\omega_2}^2} \]  

(11)

where \( \sigma_{\omega_i}^2 \) are the respective standard errors of the trends \( \omega_i \) with \( i = 1, 2 \).

The \( t \) statistic is then given by

\[ t_0 = \frac{|d|}{\sigma_d}. \]  

(12)

Accordingly the \( t \) distribution with

\[ f = \frac{(\sigma_{\omega_1}^2 + \sigma_{\omega_2}^2)^2}{(\sigma_{\omega_1}^2)^2/(l_1 - 3) + (\sigma_{\omega_2}^2)^2/(l_2 - 2)} \]  

(13)

degrees of freedom [Eq. (13) is called the Welch–Satterthwaite equation (Satterthwaite 1946)] has to be integrated from \( t_0 \) to \( \infty \). The result has to be multiplied by the factor 2 because no prior information on the sign of \( d \) exists, which requires a two-tailed test. Finally, the exceedance probability \( P(t > t_0 | H_0) \) is derived where \( t_0 \) is the result of Eq. (12) calculated from the data. The integrals of the \( t \) distribution are typically tabulated in several high level programming languages such as Octave (http://www.gnu.org/software/octave/).

c. Bayesian model intercomparison

In the following, a Bayesian method to compare trends in time series is presented. The Bayesian model selection for the difference of trends is based on the works of Bretthorst (1993) and Sivia and Skilling (2006) who estimate the difference of means and standard deviations between two sets of data. For this study, the methods have been extended to compare trends.

We set up two hypotheses:

\[ A: \text{Both sets have a common (unknown) trend } \omega; \]

\[ B: \text{The two datasets have individual (unknown) trends } \omega_1 \text{ and } \omega_2. \]

Note that the magnitudes of the trends do not matter. Hypothesis \( A \) corresponds to Eq. (7), while hypothesis \( B \) corresponds to Eqs. (5) and (6).

The posterior probability of the hypothesis \( A \), given the respective data using the Bayes theorem, is estimated:

\[ P(A|D_1, D_2, I) = \frac{P(D_1, D_2|A, I)P(A|I)}{P(D_1, D_2|I)}, \]  

(14)

where the \( D_1 \) and \( D_2 \) represent the two datasets, and \( I \) describes certain relevant background information.

Since the absolute magnitudes of the parameters \( p_1 = (\mu_{p_1}, \mu_{p_2}, \omega_p, \delta_p, \sigma_{p_1}, \sigma_{p_2}) \) from Eq. (7) are irrelevant, we can use the marginalization rule (cf. Sivia and Skilling 2006 and Bretthorst 1993) and integrate

\[ P(A|D_1, D_2, I) = \frac{\int dp_1 P(D_1, D_2|A, p_1, I)P(A, p_1|I)}{P(D_1, D_2|I)}. \]  

(15)

Assuming logical independence of the prior probabilities of the hypothesis \( A \) and the parameters \( p \),

\[ P(A, p_1|I) = P(A|I)P(p_1|I) \]  

(16)

\[ = P(A|I)P(\mu_{p_1}|I)P(\mu_{p_2}|I)P(\omega_p|I) \]  

(17)

\[ \times P(\delta_p|I)P(\sigma_{p_1}|I)P(\sigma_{p_2}|I). \]  

(18)

In the same way as in Eq. (15) the posterior for hypothesis \( B \) is derived:

\[ P(B|D_1, D_2, I) = \frac{\int dp_2 P(D_1, D_2|B, p_2, I)P(B, p_2|I)}{P(D_1, D_2|I)}, \]  

(19)

with \( p_2 = [\mu_1, \mu_2, \omega_1, \omega_2, \delta_1, \sigma_1, \sigma_2] \) from Eqs. (5) and (6) and

\[ P(B, p_2|I) = P(B|I)P(p_2|I) \]  

(20)

\[ = P(B|I)P(\mu_1|I)P(\mu_2|I)P(\omega_1|I) \]  

(21)

\[ \times P(\omega_2|I)P(\delta_1|I)P(\sigma_1|I)P(\sigma_2|I). \]  

(22)

The denominator \( P(D_1, D_2|I) \) is the evidence (cf. section 3), in this case:

\[ P(D_1, D_2|I) = P(D_1, D_2|A, I)P(A|I) \]  

(23)

\[ + P(D_1, D_2|B, I)P(B|I). \]

Here \( P(A|I) \) is the prior probability for hypothesis \( A \). As there is no reason to prefer either this hypothesis or the alternative \( P(B|I) \), we assign both with the probability 0.5; thus they cancel out in the ratios given in Eqs. (15) and (19).

The prior probabilities \( P(p_1|I) \) and \( P(p_2|I) \) in (15) and (19) do not have to be integrated because they are independent from the parameters themselves and are realized by choosing them as bounded priors in the form of fully normalized uniform distributions:

\[ P(p_i|I) = \begin{cases} 1/(p_{i,\text{max}} - p_{i,\text{min}}), & p_{i,\text{min}} < p_i < p_{i,\text{max}} \quad i = 1, 2, \\ 0, & \text{otherwise} \end{cases} \]  

(24)

That is, it is assumed that all \( p_i \) in the interval \([p_{i,\text{min}}, p_{i,\text{max}}]\) have the same probability. All prior probabilities, except
the trend priors, occur in the numerator and denominator of Eqs. (15) and (19), respectively; thus they cancel out. The priors of the pooled trend and the separate trends are chosen as $P(\omega|I) = P(\omega_p|I) = P(\omega_1|I) = P(\omega_2|I)$ with

$$P(\omega|I) = \begin{cases} 1/(\omega_{\text{max}} - \omega_{\text{min}}), & \omega_{\text{min}} \leq \omega \leq \omega_{\text{max}} \\ 0, & \text{otherwise} \end{cases} \quad (25)$$

This prior information provides the probabilistic analysis and interpretation of the results. If the range of possible trends is increased, larger differences of trends are probable and vice versa. A sensitivity analysis on the trend priors is given in section 4d. Fortunately, the trend study of Mieruch et al. (2008) provides beneficial information on the range of the trends. The boundaries for the three trend priors in Eq. (25) are chosen as $\omega_{\text{min}} = -0.1$ and $\omega_{\text{max}} = +0.1$ g cm$^{-2}$ yr$^{-1}$. This trend range comprises more than 99.9% of all vapor trends for the time span 1996 to 2006 and is a lower bound. Any decrease of the trend range would result in a truncation of the probability space.

Finally, another trend prior cancels out in each of the ratios of Eqs. (15) and (19).

The only remaining quantities are the two likelihood functions, where the two independent datasets $D_1$ and $D_2$ from radiosonde and satellite measurements are assumed to be independent on a noise basis as well:

$$P(D_1, D_2|A, \mu_1, I) = P(D_1|A, \mu_1, \omega_p, \delta_1, \sigma_1, I) \times P(D_2|A, \mu_2, \omega_p, \sigma_2, I) \quad (26)$$

and

$$P(D_1, D_2|B, \mu_2, I) = P(D_1|B, \mu_2, \omega_1, \delta_2, \sigma_1, I) \times P(D_2|B, \mu_2, \omega_2, \sigma_2, I). \quad (27)$$

A Gaussian likelihood is assumed such that the residuals $e_{1t}, e_{2t}$, and $e_{pt}$ are normally distributed. The $D_1$ comprise $l_1$ independent measurements $\{D_{1t}\}$, and the $D_2$ comprise $l_2$ independent measurements $\{D_{2t}\}$, leading to

$$P(D_1, D_2|A, \mu_1, \omega_p, \delta_1, \sigma_1, I) \propto \exp \left[ -\frac{1}{2\sigma^2_{\mu_1} \sum_{t=1}^{T_1} (D_{1t} - \mu_1 C_{1t} - \omega_p X_{1t} - \delta_1 U_t)^2} \right]$$

$$\times \exp \left[ -\frac{1}{2\sigma^2_{\omega_p} \sum_{t=1}^{T_2} (D_{2t} - \mu_2 C_{2t} - \omega_p X_{2t})^2} \right] \quad (28)$$

and

$$P(D_1, D_2|B, \mu_2, \omega_1, \delta_2, \sigma_1, I) \propto \exp \left[ -\frac{1}{2\sigma^2_{\mu_2} \sum_{t=1}^{T_1} (D_{1t} - \mu_2 C_{1t} - \omega_1 X_{1t} - \delta_2 U_t)^2} \right]$$

$$\times \exp \left[ -\frac{1}{2\sigma^2_{\omega_1} \sum_{t=1}^{T_2} (D_{2t} - \mu_2 C_{2t} - \omega_1 X_{2t})^2} \right], \quad (29)$$

where the asterisks have been dropped for convenience, but the transformed data remain addressed (cf. section 4a).

Owing to the equality of the priors and the normalization, the analysis simplifies to

$$P(A|D_1, D_2, I) = \frac{\int dp_1 P(D_1, D_2|A, \mu_1, I)}{P(D_1, D_2|I)} \quad (30)$$

and

$$P(B|D_1, D_2, I) = \frac{P(\omega|I) \int dp_2 P(D_1, D_2|B, \mu_2, I)}{P(D_1, D_2|I)}. \quad (31)$$

The final posterior probabilities (30) and (31) constitute highly complex functions comprising products of more than 200 Gaussians, which have to be integrated over six and seven dimensions, respectively. Such multidimensional, complex probability density distributions have extremely small peaks and are exceedingly steep, comparable to “needles in a haystack,” as stated by Liu (2003). This means that standard quadrature, and even standard Monte Carlo, integration algorithms are not sufficient to solve these integrals. Therefore a Markov Chain Monte Carlo (MCMC) method is used for integration [comprehensive information can be found e.g., in Gilks et al. (1995) and Robert and Casella (2005)], where the algorithm Differential Evolution Markov
Chain (DEMC), explained by ter Braak (2006), has been implemented. Two factors essentially determine the precision of the method, which are the burn in phase (bip), that is, the time the algorithm needs to converge to the target distribution, and the number of samples (nos). These parameters are adjusted to achieve a precision of $\sim 0.01$ with bip $= 10^5$ and nos $= 10^5$.

d. Sensitivity analysis

An important step in Bayesian analysis is the choice of prior information. As shown above, we have chosen the prior for the trend parameter as a fully normalized uniform distribution in the range from $\omega_{\text{min}} = -0.1$ to $\omega_{\text{max}} = +0.1$ g cm$^{-2}$ yr$^{-1}$, hence the prior range is $\Delta \omega = 0.2$ and $P(\omega|I) = \frac{1}{\Delta \omega} = 5$, which acts as a penalty on hypothesis $B$. This circumstance is known as Ockham’s Razor, a principle that recommends the selection of an accurate theory or model having the fewest assumptions and postulates when multiple competing theories are equal in describing respective phenomena. Ockham’s Razor is naturally implemented in the Bayesian concept in such a way that a theory is penalized for every additional parameter automatically.

We can qualitatively derive the Ockham factor, which is also shown in Sivia and Skilling (2006) and Dose and Menzel (2004). If model $B$ is the more complex hypothesis and model $A$ is the simpler one, there is one more dimension to integrate over, denoted as $\omega_2$, in the case of model $B$. This contribution to the integral is proportional to the width of the probability density function $P(B|D_1, D_2, I)$ in this direction (denoted as $\delta \omega_2$). With $P(\omega_2|I) = 1/\Delta \omega_2$, we see that the Ockham factor is $\approx \delta \omega_2/\Delta \omega_2$. This ratio is typically smaller than unity and thus penalizes model $B$ for its additional parameter.

On the one hand, prior information is a great advantage of Bayesian methods; however, on the other hand, the prior information represents a certain degree of subjectivity influencing the results. A sensitivity analysis with respect to the trend prior has been undertaken. Figure 1 depicts the results, where we have chosen an exemplary situation in which the probability for hypothesis $A$ yields $P(A|D_1, D_2, I) = 0.7$ and the probability for hypothesis $B$ yields $P(B|D_1, D_2, I) = 0.3$ using our chosen trend prior. These results are depicted in Fig. 1 as black and gray points. The $x$ axis in Fig. 1 represents a change of the prior probability $P(\omega|I)$ by $5$ in percent. For instance, a change of $10\%$ corresponds to an increase of the prior range of $\omega_{\text{min}} = -0.09$ and $\omega_{\text{max}} = +0.09$ g cm$^{-2}$ yr$^{-1}$. As can be seen from Fig. 1, $P(A|D_1, D_2, I)$ decreases with increasing $P(\omega|I)$ and decreasing $\Delta \omega$, and vice versa for $P(B|D_1, D_2, I)$. As mentioned in section 4c, the choice of our trend range constitutes a lower bound; thus in principle only an increase of the trend range would make sense, avoiding any truncation of the probability space. This would decrease $P(\omega|I)$ and, as can be seen in Fig. 1, increase $P(A|D_1, D_2, I)$—namely, the probability of a common trend. However, in a meaningful range of deviations from our prior, within $\pm 20\%$, the results are quite insensitive. For unrealistically large deviations of the prior from our choice, large changes of the results are observed. This is shown in the small embedded figure of Fig. 1. In conclusion, a reasonable prior has been selected, and our results are insensitive to changes of this prior—at least in a range of $\pm 20\%$.

e. Analytical approximation

DEMC is a sophisticated and powerful algorithm that goes beyond what is implemented in standard computational programming languages or packages. The disadvantage of DEMC is the need for large computational power. Sivia and Skilling (2006) have derived an approximation for a Bayesian method, which compares means and standard deviations of data. This approximation is adapted to the method for trend comparison shown in the following.

Using a quadratic Taylor series expansion of the logarithmic likelihood function in (26) we find

$$L_A = L_A(\hat{p}_1) - \frac{1}{2} K_A H_A K_A + \cdots,$$

where $L_A = \log [P(D_1|A, \mu_{p_1}, \sigma_{p_1}, \delta_p, \sigma_{p_2}, \sigma_{p_2}, I)]$ with a maximum at $\hat{p}_1 = \left[ \mu_{p_1}, \mu_{p_2}, \sigma_{p_1}, \delta_p, \sigma_{p_2} \right]$. The parameters $\hat{p}_1$ are determined by the first partial derivatives $\partial L_A/\partial \hat{p}_1 = 0$, solving a set of linear equations in a least squares sense. Thus, we can use the parameters $\hat{p}_1$ estimated in section 4a.
The second term in Eq. (32) contains the vector
\[ \mathbf{K}_A = \mathbf{p}_1 - \mathbf{p}_2, \]
which is shown explicitly in the appendix. The entries of the \( 6 \times 6 \) matrix \( \mathbf{H}_A \) in Eq. (32) are derived from the second partial derivatives of \( L_A \) evaluated at \( \mathbf{p}_1 \), which is shown in the appendix.

The approximated likelihood of hypothesis \( A \) exponentiating \( L_A \), is
\[
P_{\text{approx}}(\mathbf{D}_1, \mathbf{D}_2|A, I) = \int d\mathbf{p}_1 \exp L_A
\]
\[
= \int d\mathbf{p}_1 \exp [L_A(\mathbf{p}_1)]
\times \exp \left[ -\frac{1}{2} \mathbf{K}_A' \mathbf{H}_A \mathbf{K}_A \right].
\]

The first exponential in Eq. (34) is a constant and the second is a six-dimensional Gaussian. Sivia and Skilling (2006) integrate an \( M \)-dimensional Gaussian by,
\[
Z = \int \exp \left[ -\frac{1}{2} \mathbf{x}' \mathbf{H} \mathbf{x} \right] d^M \mathbf{x}
= \frac{(2\pi)^{M/2}}{\sqrt{\det \mathbf{H}}},
\]
thus Eq. (34) becomes
\[
P_{\text{approx}}(\mathbf{D}_1, \mathbf{D}_2|A, I) = \exp[L_A(\mathbf{p}_1)] \frac{(2\pi)^{6/2}}{\sqrt{\det \mathbf{H}_A}}.
\]
This analysis yields
\[
P_{\text{approx}}(\mathbf{D}_1, \mathbf{D}_2|A, I) = (\hat{\sigma}_p)^{-1} \frac{(2\pi)^{6/2}}{\sqrt{\det \mathbf{H}_A}}.
\]

The alternative hypothesis \( B \), stating that the time series have individual trends \( \omega_1 \) and \( \omega_2 \), is now derived. The procedure is identical to the previous derivations, using the quadratic Taylor series expansion of the logarithmic likelihood function Eq. (27)
\[
L_B = L_B(\mathbf{p}_2) - \frac{1}{2} \mathbf{K}_B' \mathbf{H}_B \mathbf{K}_B + \cdots.
\]
The quantities \( \mathbf{p}_2, \mathbf{K}_B \), and \( \mathbf{H}_B \) are given in the appendix. Accordingly, we find
\[
P_{\text{approx}}(\mathbf{D}_1, \mathbf{D}_2|B, I) = (\hat{\sigma}_p)^{-1} \frac{(2\pi)^{3/2}}{\sqrt{\det \mathbf{H}_B}}.
\]
The posterior probabilities have to be normalized. Because of
\[
P_{\text{approx}}(A|D_1, D_2, I) + P_{\text{approx}}(B|D_1, D_2, I) = 1
\]
with the evidence
\[
P_{\text{approx}}(D_1, D_2|I) = P_{\text{approx}}(D_1, D_2|A, I)
+ P_{\text{approx}}(D_1, D_2|B, I),
\]
this yields
\[
P_{\text{approx}}(A|D_1, D_2, I) = \frac{P_{\text{approx}}(D_1, D_2|A, I)}{P_{\text{approx}}(D_1, D_2|B, I)}.
\]
and
\[
P_{\text{approx}}(B|D_1, D_2, I) = \frac{P_{\text{approx}}(D_1, D_2|B, I)P(\omega|I)}{P_{\text{approx}}(D_1, D_2|I)}.
\]
After normalization all terms occurring in both (42) and (43) cancel out in the ratios, thus we have
\[
P_{\text{approx}}(D_1, D_2|A, I) = \frac{(2\pi)^{6/2}}{\hat{\sigma}_1^4 \hat{\sigma}_2^4 \sqrt{\det \mathbf{H}_A}}.
\]
and
\[
P_{\text{approx}}(D_1, D_2|B, I) = \frac{(2\pi)^{3/2}}{\hat{\sigma}_1^2 \hat{\sigma}_2^2 \sqrt{\det \mathbf{H}_B}}.
\]
The posterior probability of hypothesis \( B \) is also proportional to the prior probability of the trends, which is chosen in the same way as in Eq. (25). Eqs. (42) and (43) are analytical functions, which can quite easily be computed in contrast to Eqs. (30) and (31) that can only be solved numerically.

5. Results

The comparison methods for trends in time series described above, that is, the Welch test and the Bayesian model selection, have been applied to measured trends from satellite and radiosonde monthly mean water vapor data. The trends have been calculated using the methods described in section 4a. For the comparison a quality criterion is required, that is, both time series have to contain at least two-thirds of the monthly mean measurements over the time span from January 1996 to December 2007, that is, at least 96 data points from the
maximal 144. This constraint assures that the trends are representative for the period investigated and less susceptible to possible outliers.

Figure 2a depicts the results of the Welch test. The probabilities $P(t > t_0|H_0)$, with null hypothesis $H_0$: $d = \omega_1 - \omega_2 = 0$ and the difference of the trends $d$, for the 187 trend pairs are plotted versus $d$. High probabilities are observed for small trend differences, while lower probabilities are found for large trend differences, as expected. Figure 2b shows the $P(t > t_0|H_0)$ plotted...
versus the trend differences normalized to the error of the difference. From the definition of the Welch test it is clear that \( P(t > t_0|H_0) \) is totally determined by \( (\omega_1 - \omega_2)/\sigma_d \), where \( \omega_1 \) is the GOME–SCIAMACHY trend and \( \omega_2 \) is the radiosonde trend. In the sense of the frequentist interpretation, the null hypothesis for a single test would be rejected if \( P(t > t_0|H_0) < \alpha = 0.05 \), which would apply in 20 of 187 cases (\( \sim 10\% \)). Hence, in about 90% of the tests we cannot reject the null hypothesis. However, we are dealing with 187 independent tests, so multiple testing would be possibly more appropriate. A common approach used in multiple comparisons is the Bonferroni correction (see, e.g., Perneger 1998), which decreases the significance level \( \alpha \). The reason for this correction is that the number of observed significant test results that occur by chance, performing \( n \) tests, is \( \leq na \). The Bonferroni correction would yield a significance level: 
\[
\tilde{\alpha} = 1 - (1 - \alpha)^{1/n} = 1 - (1 - 0.05)^{1/187} = 0.0027
\]
Using the Bonferroni correction we would test the general null hypothesis that all null hypotheses are true simultaneously, which has to be rejected if one or more \( p \) values are smaller than the significance level \( \tilde{\alpha} \). Since no \( P(t > t_0|H_0) \) from the 187 tests is smaller than \( \tilde{\alpha} \), the general null hypothesis cannot be rejected. The Bonferroni correction is quite conservative, which has been seriously criticized, for example, by Perneger (1998). A discussion on the usefulness of the Bonferroni correction is considered valuable, but the general conclusions of this investigation are similar with and without the Bonferroni correction. Without the multiple testing correction the null hypothesis is accepted in 90% of all cases. Using the Bonferroni correction the general null hypothesis for the complete dataset is also accepted.

The 187 probabilities of a common trend \( P(A|D_1, D_2, I) \) from the exact Bayesian model selection (for each trend pair) are plotted versus the difference of the trends in Fig. 2c and versus the trend difference normalized to the error of the difference \( (\omega_1 - \omega_2)/\sigma_d \) in Fig. 2d. Additionally, the results from the approximation of the Bayesian method are shown in Figs. 2c and 2f. High probabilities for a common trend are found for small trend differences, whereas the probability is low for large trend deviations as in the case of the Welch test. The approximation slightly overestimates the exact probabilities and the mean relative difference is \( O(10\%) \), but the general results from the exact method and the approximation are very similar, thus the use of the approximation is recommended for monthly mean water-vapor trend comparison if sophisticated algorithms like DEMC are not available or large datasets have to be analyzed in a short period of time. To judge the probabilities of different hypotheses, Jeffreys (1939) introduced the scale presented in Table 1. Here \( P(A) \) and \( P(B) \) denote the respective probabilities of the hypotheses, where a value of \( \log_{10}[P(A)/P(B)] = 1 \) means that hypothesis A is 10 times more probable than hypothesis B.

Regarding the Jeffreys scale (Table 1), the evidence against hypothesis B is substantial if the logarithm of the so-called Bayes factor, which is \( P(D_1, D_2|A, I)/[P(D_1, D_2|B, I) \cdot P(w|I)] \), is larger than 0.5 and smaller than 1, which corresponds to \( 0.76 < P(A|D_1, D_2, I) < 0.91 \); hence the evidence against hypothesis A is substantial if \( 0.09 < P(A|D_1, D_2, I) < 0.24 \). Hypothesis A is preferred substantially in 49 cases and hypothesis B in 9 cases, using the exact method. When the approximation is used, A is preferred in 114 cases and B in 5 cases. The evidence against B is strong to decisive if \( P(A|D_1, D_2, I) > 0.91 \), which is true in zero cases for the exact solution and true in 10 cases for the approximation. Strong to decisive evidence is drawn against A if \( P(A|D_1, D_2, I) < 0.09 \), which is observed 3 times in the exact case and 2 times in the case of the approximation. The rigorous application of the Bayesian model selection would prefer hypothesis A if \( P(A|D_1, D_2, I) > 0.5 \), which is true in 153 cases of 187, that is, 82% for the exact method and in 156 cases for the approximation. Interpreting the observed patterns in Figs. 2c–f, distinct clusters of data points are found between probabilities of 0.7 to 0.9. These are mostly classified as substantially supporting hypothesis A of a common trend. Since this is true for the exact Bayesian method and the approximation, generally similar conclusions are drawn. Nevertheless, since large differences between the exact and the approximation method have been observed for single time series, significant conclusions can only be drawn for the ensemble level. Furthermore, as applied for the Welch test, the argumentation of multiple testing also concerns the Bayesian model selection. Westfall et al. (1997) (and citations therein) suggest a perspective of a Bayesian Bonferroni correction, which acts on the prior information. Accordingly, our trend prior \( P(w|I) \) would be transformed to \( P(w|I) = P(w|I)^{1/0.2} = (1/0.2)^{1/0.2} = (1/0.2)^{1/187} \approx 1 \), which corresponds to an enlarged trend range of \( \omega_{\text{min}} = -0.5 \) and \( \omega_{\text{max}} = +0.5 \text{ g cm}^{-2} \text{ yr}^{-1} \), increasing

<table>
<thead>
<tr>
<th>( \log_{10}[P(A)/P(B)] )</th>
<th>Evidence against B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 to 0.5</td>
<td>Nothing but a chance</td>
</tr>
<tr>
<td>0.5 to 1</td>
<td>Substantial</td>
</tr>
<tr>
<td>1 to 2</td>
<td>Strong</td>
</tr>
<tr>
<td>&gt;2</td>
<td>Decisive</td>
</tr>
</tbody>
</table>
the $P(A|D_1, D_2, I) = 0.82$. The approximation method gives $P_{\text{approx}}(A|D_1, D_2, I) = 0.90$. Hence, for small trend differences, both the frequentist and Bayesian concept reveal quite large probabilities for the respective tests.

Low probabilities are found, for example, at Albany Airport in Australia. The time series are shown in Figs. 4c and 4d. The visual inspection definitely classifies the trends as different. The trend difference is actually $\omega_1 - \omega_2 = 0.04$ g cm$^{-2}$ yr$^{-1}$ and after normalization it is $(\omega_1 - \omega_2)/\sigma_d = 3.2$. The Welch test gives a probability of $P(t > t_0|H_0) = 0.002$ (rejection of the null hypothesis).

The exact Bayesian finds $P(A|D_1, D_2, I) = 0.02$ and the approximation yields $P_{\text{approx}}(A|D_1, D_2, I) = 0.04$ (preferring $B$). Thus, low probabilities are found for large trend differences by both statistical methods. Different probabilities for the frequentist test and the Bayesian method, as can be seen from Fig. 2, are found in the range between small and large trend differences. As an example, a pair of water vapor time series from Meiningen, Germany, is chosen with a trend difference of $\omega_1 - \omega_2 = 0.014$ g cm$^{-2}$ yr$^{-1}$ and a normalized trend difference of $(\omega_1 - \omega_2)/\sigma_d = 1.3$. The probability derived from the Welch test yields $P(t > t_0|H_0) = 0.19$, which is small, but not small enough to reject the null hypothesis. The exact Bayesian gives $P(A|D_1, D_2, I) = 0.89$, and the approximation is $P_{\text{approx}}(A|D_1, D_2, I) = 0.75$. Here it again has to be mentioned that both methods (Welch test/Bayes) reveal different probabilities and are both correct under the respective frameworks of the frequentist philosophy and the Bayesian concept. The exact probabilities of the frequentist and Bayesian method differ; however, the conclusions are nevertheless similar.

The GOME–SCIAMACHY trends are plotted in Fig. 5, where the 187 radiosonde trends have been embedded into the figure indicated by black, gray, and white bordered circles. The circles of radiosonde trends are filled with the color for the magnitude of the respective trends according to the color bar used for the GOME–SCIAMACHY data as well. The borders of the circles indicate the Bayesian posterior probabilities $P(A|D_1, D_2, I)$ at a specific geolocation. A black border indicates a probability $\leq 0.5$, which means that hypothesis $B$ is preferred. Seven of the total 34 black bordered circles are covered by the other circles and cannot be seen in the figure. A gray-bordered circle represents probabilities $>0.5$ and $\leq 0.76$ where hypothesis $A$ is favored (104 circles), and a white border indicates that hypothesis $A$ is substantially preferred with Bayesian probabilities $>0.76$ (49 circles).

One reason for discrepancies between satellite and radiosonde trends are data gaps in the radiosonde data. This has been observed, for example, at Minqin, China, shown in Figs. 4g and 4h. Radiosonde data are often missing in summer, especially in 2006 and 2007 when high water vapor was observed by SCIAMACHY. The Welch test gives $P(t > t_0|H_0) = 0.01$, which implies that the null hypothesis should be rejected. The exact Bayesian gives $P(A|D_1, D_2, I) = 0.43$, hence preferring hypothesis $B$. Again, the individual $p$ values of both methods are different, but the conclusions are similar.

As mentioned in section 2b one possible important reason for discrepancies between observed trends from satellite and radiosonde water vapor data is the different resolution of the two instruments. Radiosondes can capture local events, whereas the satellite measurement is an average over a large area. This will be shown in the following using an example from the west coast of Saudi Arabia. A zoom into this region is depicted in Fig. 6. Note that in Fig. 6 the color scale used for the
Fig. 4. Examples of GOME–SCIAMACHY and radiosonde water vapor time series at (a),(b) Nottingham, England; (c),(d) Albany Airport, Australia; (e),(f) Meiningen, Germany; and Minqin, China.

(a) GOME/SCIAMACHY deseasonalized data (black) and radiosonde data (gray).

(b) Pooled GOME/SCIAMACHY and radiosonde water vapor data.

(c) GOME/SCIAMACHY deseasonalized data (black) and radiosonde data (gray).

(d) Pooled GOME/SCIAMACHY and radiosonde water vapor data.

(e) GOME/SCIAMACHY deseasonalized data (black) and radiosonde data (gray).

(f) Pooled GOME/SCIAMACHY and radiosonde water vapor data.

(g) GOME/SCIAMACHY deseasonalized data (black) and radiosonde data (gray).

(h) Pooled GOME/SCIAMACHY and radiosonde water vapor data.

Fig. 4. Examples of GOME–SCIAMACHY and radiosonde water vapor time series at (a),(b) Nottingham, England, (c),(d) Albany Airport, Australia; (e),(f) Meiningen, Germany; and Minqin, China.
GOME–SCIAMACHY and radiosonde trends is different from the one used in Fig. 5. Here a positive water vapor trend is observed with a radiosonde measurement located exactly at the city of Jeddah. The satellite trends in the near vicinity of the town are enhanced as well but are not as strong as the very localized radiosonde trend. However, it seems possible that changes in the total water vapor column, observed via satellite, can be attributed to human activities with a high probability, identifying Jeddah as a source of water vapor. Further, increasing water vapor is observed exactly at the city of Asmara in Eritrea as well (cf. Fig. 6), which shows that urban areas, and hence anthropogenic influence on water vapor changes, can be detected using satellite observations. The satellite measurements of the positive trends over Jeddah are rather smeared out over a larger region. This is a likely explanation for the relatively low probability for hypothesis \( A \) between the observed trends, which is indicated by the gray-bordered circle.

6. Conclusions

In this manuscript, we have solved the Behrens–Fisher problem (same means, unequal unknown standard deviations) for the analysis of trends from temporal series for water vapor, derived by different methods using two schools of statistics, the frequentist (standard) approach and the Bayesian concepts. We have applied both methods to global water vapor datasets observed by satellite (GOME–SCIAMACHY) and radiosonde instruments.

To utilize and assess the value of frequentist statistics, the widely used Welch test has been applied to address the Behrens–Fisher problem. The individual null hypothesis \( H_0: d = \omega_1 - \omega_2 = 0 \), stating that the difference of the trends is equal to zero, could only be rejected in 10% of the 187 trend comparisons. Additionally we have applied multiple testing by using the Bonferroni correction to test the general null hypothesis stating that all single null hypotheses are true, which could not be rejected. These results increase our confidence, in a climatological sense, that satellite and radiosonde measures the true changes of total column water vapor, assuming no long-term calibration issues for the radiosondes, negligible changes in cloud cover, and an adequate combination of the GOME and SCIAMACHY data.
Concerning the Bayesian model selection we estimated the probabilities for the hypotheses

A: Both sets have a common (unknown) trend \( \omega \);
B: The two datasets have individual (unknown) trends \( \omega_1 \) and \( \omega_2 \).

This is achieved using prior information and further the complete data under consideration. Today, the computational power of standard computers is sufficient to operate, for example, the genetic algorithm DEMC to estimate the Bayesian posteriors. In the Bayesian framework, a cluster of about 50 trend differences between probabilities of 0.7 and 0.9 is obtained for hypothesis A (cf. Fig. 2). This provides evidence supporting the hypothesis A, namely, that the observed trends are common for the two datasets. Overall, the analysis of the datasets for the two different hypotheses of having either common and individual trends show that the common trend is preferred in 153 cases of 187 (82%) and that generally hypothesis A has a larger probability than hypothesis B.

We attribute the cases where the results are poorer to either an inadequate calibration of the long-term datasets from the radiosondes or local-scale changes, which are smoothed in the lower spatially resolved remote sounding dataset. In addition, the use of a Bonferroni-like correction in a Bayesian sense results in a clear preference of hypothesis A, thus hinting at a common trend in satellite and radiosonde measurements.

In conclusion both approaches, the frequentist Welch test and the Bayesian model selection, yield similar results when used to compare water vapor trends from the two independent measurement sets on an ensemble basis, including or neglecting multiple testing. However, differences can occur when using the two tests to compare trends at a given location. The frequentist method is recommended, provided it is sufficient to work with a given null hypothesis and thereby infer the probability of a parameter and associated significance test. If this is not the case, then we recommend the Bayesian approach, where it is possible to derive probabilities of competing hypotheses. From a climatological viewpoint evidence that trends in time series from both observing systems, radiosondes, and satellites are real significant and not statistical artifacts has been found. The size of the trends is not discussed further because the length of the time series is short and therefore the trends observed are not yet identifying a climate signal.

Acknowledgments. SCIAMACHY is a national contribution to the ESA ENVISAT project, funded by Germany, the Netherlands, and Belgium. SCIAMACHY data have been provided by ESA. Radiosonde data have been provided by DWD. This work has been funded in part by the University and state of Bremen, the German Aerospace (DLR) by the EU (6th and 7th framework ACCENT and ACCENT plus projects), and by EUMETSAT in the framework of the Climate Monitoring (CM-SAF) part of the Satellite Application Facilities Network. We thank D. S. Sivia and G. L. Bretthorst for advice concerning prior probabilities. We also thank C. ter Braak for discussion on the DEMC algorithm. Fruitful discussions are acknowledged to A. C. Davison and J. A. Freund during the 11th IMSC in Edinburgh. We thank Kirsten Schnülle for proofreading the manuscript. We grateful acknowledge the critical comments of three anonymous referees, which have improved the manuscript significantly.

APPENDIX

Analytical Approximation—The Matrices

Regarding section 4e, the quadratic Taylor series expansion of the logarithmic likelihood function Eq. (26) yields

\[
L_A = L_A(\hat{p}_1) - \frac{1}{2} K'_A H_A K_A + \cdots, \tag{A1}
\]

with

\[
K'_A = p_1 - \hat{p}_1 = (\mu_{p_1} - \hat{\mu}_{p_1}, \mu_{p_2} - \hat{\mu}_{p_2}, \omega_p - \hat{\omega}_p, \delta_p - \hat{\delta}_p, \alpha_{p_1} - \hat{\alpha}_{p_1}, \alpha_{p_2} - \hat{\alpha}_{p_2}) \tag{A2}
\]

and

Fig. 6. GOME–SCIAMACHY water vapor trends in the range of \( \pm 0.03 \, \text{g cm}^{-2} \, \text{yr}^{-1} \) at the Arabian Peninsula (10°–30°N, 35°–60°E) with embedded radiosonde trends from 1996 to 2007. A different color scale as in Fig. 5 is used in this blowup.
with

$$K_B' = p_2 - \bar{p}_2 = (\mu_1 - \mu_1 \bar{\mu}_2 - \bar{\mu}_2 \omega_1 - \bar{\omega}_1 \omega_2 - \bar{\omega}_2 \delta - \delta \bar{\sigma}_1 - \bar{\sigma}_1 \sigma_2 - \bar{\sigma}_2)$$

and

Analogously the quadratic Taylor series expansion of the logarithmic likelihood function Eq. (27) yields

$$L_B = L_B(\bar{p}_2) - \frac{1}{2} K_B' H_B K_B + \cdots,$$

(A5)
\[
H_B = \begin{pmatrix}
\frac{\partial^2 L_B}{\partial \mu_1^2} & \frac{\partial^2 L_B}{\partial \mu_1 \partial \mu_2} & \frac{\partial^2 L_B}{\partial \mu_1 \partial \omega_1} & \frac{\partial^2 L_B}{\partial \mu_1 \partial \omega_2} & \frac{\partial^2 L_B}{\partial \mu_1 \partial \delta} & \frac{\partial^2 L_B}{\partial \mu_1 \partial \sigma_1} & \frac{\partial^2 L_B}{\partial \mu_1 \partial \sigma_2} \\
\frac{\partial^2 L_B}{\partial \mu_2^2} & \frac{\partial^2 L_B}{\partial \mu_2 \partial \mu_1} & \frac{\partial^2 L_B}{\partial \mu_2 \partial \omega_1} & \frac{\partial^2 L_B}{\partial \mu_2 \partial \omega_2} & \frac{\partial^2 L_B}{\partial \mu_2 \partial \delta} & \frac{\partial^2 L_B}{\partial \mu_2 \partial \sigma_1} & \frac{\partial^2 L_B}{\partial \mu_2 \partial \sigma_2} \\
\frac{\partial^2 L_B}{\partial \omega_1^2} & \frac{\partial^2 L_B}{\partial \omega_1 \partial \mu_1} & \frac{\partial^2 L_B}{\partial \omega_1 \partial \omega_2} & \frac{\partial^2 L_B}{\partial \omega_1 \partial \delta} & \frac{\partial^2 L_B}{\partial \omega_1 \partial \sigma_1} & \frac{\partial^2 L_B}{\partial \omega_1 \partial \sigma_2} \\
\frac{\partial^2 L_B}{\partial \omega_2^2} & \frac{\partial^2 L_B}{\partial \omega_2 \partial \mu_1} & \frac{\partial^2 L_B}{\partial \omega_2 \partial \omega_1} & \frac{\partial^2 L_B}{\partial \omega_2 \partial \delta} & \frac{\partial^2 L_B}{\partial \omega_2 \partial \sigma_1} & \frac{\partial^2 L_B}{\partial \omega_2 \partial \sigma_2} \\
\frac{\partial^2 L_B}{\partial \delta^2} & \frac{\partial^2 L_B}{\partial \delta \partial \mu_1} & \frac{\partial^2 L_B}{\partial \delta \partial \omega_1} & \frac{\partial^2 L_B}{\partial \delta \partial \omega_2} & \frac{\partial^2 L_B}{\partial \delta \partial \sigma_1} & \frac{\partial^2 L_B}{\partial \delta \partial \sigma_2} \\
\frac{\partial^2 L_B}{\partial \sigma_1^2} & \frac{\partial^2 L_B}{\partial \sigma_1 \partial \mu_1} & \frac{\partial^2 L_B}{\partial \sigma_1 \partial \omega_1} & \frac{\partial^2 L_B}{\partial \sigma_1 \partial \omega_2} & \frac{\partial^2 L_B}{\partial \sigma_1 \partial \delta} & \frac{\partial^2 L_B}{\partial \sigma_1 \partial \sigma_2} \\
\frac{\partial^2 L_B}{\partial \sigma_2^2} & \frac{\partial^2 L_B}{\partial \sigma_2 \partial \mu_1} & \frac{\partial^2 L_B}{\partial \sigma_2 \partial \omega_1} & \frac{\partial^2 L_B}{\partial \sigma_2 \partial \omega_2} & \frac{\partial^2 L_B}{\partial \sigma_2 \partial \delta} & \frac{\partial^2 L_B}{\partial \sigma_2 \partial \sigma_1}
\end{pmatrix}
\]

which is

\[
H_B = \begin{pmatrix}
\frac{l_1}{\sigma_1^2} & 0 & \frac{\Sigma X_{1u}}{\sigma_1^2} & 0 & \frac{\Sigma U_j}{\sigma_1^2} & 2 & 0 \\
0 & \frac{l_2}{\sigma_2^2} & 0 & \frac{\Sigma X_{2u}}{\sigma_2^2} & 0 & 0 & 0 & \frac{2}{\sigma_2^2} \\
\frac{\Sigma X_{1u}}{\sigma_1^2} & 0 & \frac{\Sigma X_{2u}^2}{\sigma_1^2} & 0 & \frac{\Sigma U_j X_{1u}}{\sigma_1^2} & 0 & 0 & \frac{2\Sigma \epsilon_{2u} X_{2u}}{\sigma_1^2} \\
0 & \frac{\Sigma X_{2u}}{\sigma_2^2} & 0 & \frac{\Sigma X_{2u}^2}{\sigma_2^2} & 0 & 0 & 0 & \frac{2\Sigma \epsilon_{2u} X_{2u}}{\sigma_2^2} \\
\frac{\Sigma U_j}{\sigma_1^2} & 0 & \frac{\Sigma U_j X_{1u}}{\sigma_1^2} & 0 & \frac{\Sigma U_j^2}{\sigma_1^2} & \frac{2\Sigma \epsilon_{1u} U_j}{\sigma_1^2} & 0 & 0 \\
\frac{2}{\sigma_1^2} & 0 & \frac{2\Sigma \epsilon_{1u} X_{1u}}{\sigma_1^2} & 0 & \frac{2\Sigma \epsilon_{1u}^2}{\sigma_1^2} & \frac{2\Sigma \epsilon_{1u} X_{1u}}{\sigma_1^2} & 0 & 0 \\
0 & \frac{2}{\sigma_2^2} & 0 & \frac{2\Sigma \epsilon_{2u} X_{2u}}{\sigma_2^2} & 0 & 0 & \frac{2\Sigma \epsilon_{2u} X_{2u}}{\sigma_2^2}
\end{pmatrix}
\]

REFERENCES


Welch, B. L., 1947: The generalization of “student’s” problem when several different population variances are involved. Biometrika, 34, 28–35.