UCAR | COMMUNITY PROGRAMS

COSMIC

The N-Cornered Hat method for Estimating Error Variances between Multiple Data Sets:

Theoretical considerations and comparisons with the two-cornered hat method

Jeremiah P Sjoberg*, Richard A Anthes, and Therese Rieckh

UCAR/COSMIC

Abstract

Variations of the three-cornered hat (3CH) method for estimating random error variances associated with three or more data sets have been reported in the literature for studying geophysical data sets, including sea-surface winds, sea-surface temperatures, precipitation, leaf area index and soil moisture. Anthes and Rieckh (2018) and Rieckh and Anthes (2018) used the 3CH method to estimate the error variances of multiple atmospheric sounding data sets. However the methods used often contain subtle variations and make different assumptions. Here we derive the full 3CH equations that relate the error variance of three or more observations to the variance of differences between the data sets and the error covariances among the different data sets.

Three-cornered hat relations

Consider the data sets X_1, X_2, and X_3 with individual elements $i = 1, 2, \ldots$. We assume that these may be cast as

$$X_{1,i} = T_i + b_{X1} + \varepsilon_{X1,i} \tag{1a}$$
$$X_{2,i} = T_i + b_{X2} + \varepsilon_{X2,i} \tag{1b}$$
$$X_{3,i} = T_i + b_{X3} + \varepsilon_{X3,i}\,, \tag{1c}$$

where T_i is a set of reference values; the b terms are mean differences of the individual data sets from the reference data set; and the ε terms are sets of zero mean, not necessarily Gaussian random variations.

To remove the mean difference terms b, we subtract the mean $E[\cdot]$ of each data set:

$$X'_{1,i} = T'_i + \varepsilon_{X1,i} \tag{2a}$$
$$X'_{2,i} = T'_i + \varepsilon_{X2,i} \tag{2b}$$
$$X'_{3,i} = T'_i + \varepsilon_{X3,i}\,, \tag{2c}$$

where primes denote difference from the mean.

The unique set of the variance of differences between data sets can be written

$$\begin{aligned}\mathrm{Var}\left[X_{1,i} - X_{2,i}\right] &= \mathrm{Var}\left[\varepsilon_{X1,i}\right] + \mathrm{Var}\left[\varepsilon_{X2,i}\right] \\ &\quad - 2\mathrm{Cov}\left[\varepsilon_{X1,i}, \varepsilon_{X2,i}\right] \end{aligned}\tag{3a}$$
$$\begin{aligned}\mathrm{Var}\left[X_{1,i} - X_{3,i}\right] &= \mathrm{Var}\left[\varepsilon_{X1,i}\right] + \mathrm{Var}\left[\varepsilon_{X3,i}\right] \\ &\quad - 2\mathrm{Cov}\left[\varepsilon_{X1,i}, \varepsilon_{X3,i}\right] \end{aligned}\tag{3b}$$
$$\begin{aligned}\mathrm{Var}\left[X_{2,i} - X_{3,i}\right] &= \mathrm{Var}\left[\varepsilon_{X2,i}\right] + \mathrm{Var}\left[\varepsilon_{X3,i}\right] \\ &\quad - 2\mathrm{Cov}\left[\varepsilon_{X2,i}, \varepsilon_{X3,i}\right]\,, \end{aligned}\tag{3c}$$

where $\mathrm{Cov}[\cdot]$ is the covariance between two quantities.

The relations for error variance can be derived by linearly combining Eqs. (3a-c):

$$\begin{aligned}\mathrm{Var}\left[\varepsilon_{X1,i}\right] &= \frac{1}{2}\left(\mathrm{Var}\left[X_{1,i} - X_{2,i}\right] + \mathrm{Var}\left[X_{1,i} - X_{3,i}\right] - \mathrm{Var}\left[X_{2,i} - X_{3,i}\right]\right) \\ &\quad + \mathrm{Cov}\left[\varepsilon_{X1,i}, \varepsilon_{X2,i}\right] + \mathrm{Cov}\left[\varepsilon_{X1,i}, \varepsilon_{X3,i}\right] - \mathrm{Cov}\left[\varepsilon_{X2,i}, \varepsilon_{X3,i}\right]\end{aligned}\tag{4a}$$
$$\begin{aligned}\mathrm{Var}\left[\varepsilon_{X2,i}\right] &= \frac{1}{2}\left(\mathrm{Var}\left[X_{1,i} - X_{2,i}\right] + \mathrm{Var}\left[X_{2,i} - X_{3,i}\right] - \mathrm{Var}\left[X_{1,i} - X_{3,i}\right]\right) \\ &\quad + \mathrm{Cov}\left[\varepsilon_{X1,i}, \varepsilon_{X2,i}\right] + \mathrm{Cov}\left[\varepsilon_{X2,i}, \varepsilon_{X3,i}\right] - \mathrm{Cov}\left[\varepsilon_{X1,i}, \varepsilon_{X3,i}\right]\end{aligned}\tag{4b}$$
$$\begin{aligned}\mathrm{Var}\left[\varepsilon_{X3,i}\right] &= \frac{1}{2}\left(\mathrm{Var}\left[X_{1,i} - X_{3,i}\right] + \mathrm{Var}\left[X_{2,i} - X_{3,i}\right] - \mathrm{Var}\left[X_{1,i} - X_{2,i}\right]\right) \\ &\quad + \mathrm{Cov}\left[\varepsilon_{X1,i}, \varepsilon_{X3,i}\right] + \mathrm{Cov}\left[\varepsilon_{X2,i}, \varepsilon_{X3,i}\right] - \mathrm{Cov}\left[\varepsilon_{X1,i}, \varepsilon_{X2,i}\right]\,.\end{aligned}\tag{4c}$$

Four-cornered hat relations

Following the derivation of the 3CH relations, it may be shown that the four-cornered hat error variance relation for data set $X_{1,i}$ is

$$\begin{aligned}\mathrm{Var}\left[\varepsilon_{X1,i}\right] &= \frac{1}{3}\left(\mathrm{Var}\left[X_{1,i} - X_{2,i}\right] + \mathrm{Var}\left[X_{1,i} - X_{3,i}\right] + \mathrm{Var}\left[X_{1,i} - X_{4,i}\right]\right) \\ &\quad - \frac{1}{6}\left(\mathrm{Var}\left[X_{2,i} - X_{3,i}\right] + \mathrm{Var}\left[X_{2,i} - X_{4,i}\right] + \mathrm{Var}\left[X_{3,i} - X_{4,i}\right]\right) \\ &\quad + \frac{2}{3}\left(\mathrm{Cov}\left[\varepsilon_{X1,i}, \varepsilon_{X2,i}\right] + \mathrm{Cov}\left[\varepsilon_{X1,i}, \varepsilon_{X3,i}\right] + \mathrm{Cov}\left[\varepsilon_{X1,i}, \varepsilon_{X4,i}\right]\right) \\ &\quad - \frac{1}{3}\left(\mathrm{Cov}\left[\varepsilon_{X2,i}, \varepsilon_{X3,i}\right] + \mathrm{Cov}\left[\varepsilon_{X2,i}, \varepsilon_{X4,i}\right] + \mathrm{Cov}\left[\varepsilon_{X3,i}, \varepsilon_{X4,i}\right]\right)\,.\end{aligned}\tag{5}$$

Contact*

sjoberg@ucar.edu

N-cornered hat relations

The error variance relation derivation may be generalized for N data sets such that we may write the error variance for $X_{1,i}$ as

$$\begin{aligned}\mathrm{Var}[\varepsilon_{1,i}] &= \frac{1}{N-1}\sum_{n=2}^{N}\mathrm{Var}[X_{1,i} - X_{n,i}] \\ &\quad - \frac{1}{(N-1)(N-2)}\sum_{n=2}^{N-1}\sum_{m=n+1}^{N}\mathrm{Var}[X_{n,i} - X_{m,i}] \\ &\quad + \frac{2}{N-1}\sum_{n=2}^{N}\mathrm{Cov}[X_{1,i}, X_{n,i}] \\ &\quad - \frac{2}{(N-1)(N-2)}\sum_{n=2}^{N-1}\sum_{m=j+1}^{N}\mathrm{Cov}[X_{n,i}, X_{m,i}]\,,\end{aligned}$$

and so on for the additional $N-1$ data sets.

3CH relations using 4 data sets

Note that we may use four data sets with the 3CH method to get three separate relations for the error variance of each data set. E.g., consider $X_{1,i}$ and assume all error covariance terms are 0 – a standard assumption for applying the 3CH method to real data – we find

$$\mathrm{Var}\left[\varepsilon_{X1,i}\right] = \frac{1}{2}\left(\mathrm{Var}[X_{1,i} - X_{2,i}] + \mathrm{Var}[X_{1,i} - X_{3,i}] - \mathrm{Var}[X_{2,i} - X_{3,i}]\right)$$
$$\mathrm{Var}\left[\varepsilon_{X1,i}\right] = \frac{1}{2}\left(\mathrm{Var}[X_{1,i} - X_{2,i}] + \mathrm{Var}[X_{1,i} - X_{4,i}] - \mathrm{Var}[X_{2,i} - X_{4,i}]\right)$$
$$\mathrm{Var}\left[\varepsilon_{X1,i}\right] = \frac{1}{2}\left(\mathrm{Var}[X_{1,i} - X_{3,i}] + \mathrm{Var}[X_{1,i} - X_{4,i}] - \mathrm{Var}[X_{3,i} - X_{4,i}]\right)\,.$$

Combining,

$$\begin{aligned}3\mathrm{Var}\left[\varepsilon_{X1,i}\right] &= \frac{1}{2}(\mathrm{Var}[X_{1,i} - X_{2,i}] + \mathrm{Var}[X_{1,i} - X_{3,i}] + \mathrm{Var}[X_{1,i} - X_{2,i}] \\ &\quad + \mathrm{Var}[X_{1,i} - X_{4,i}] + \mathrm{Var}[X_{1,i} - X_{3,i}] + \mathrm{Var}[X_{1,i} - X_{4,i}] \\ &\quad - \mathrm{Var}[X_{2,i} - X_{3,i}] - \mathrm{Var}[X_{2,i} - X_{4,i}] - \mathrm{Var}[X_{3,i} - X_{4,i}])\,.\end{aligned}$$

This can be reduced to

$$\begin{aligned}\mathrm{Var}[\varepsilon_{X1,i}] &= \frac{1}{3}\left(\mathrm{Var}[X_{1,i} - X_{2,i}] + \mathrm{Var}[X_{1,i} - X_{3,i}] + \mathrm{Var}[X_{1,i} - X_{4,i}]\right) \\ &\quad - \frac{1}{6}\left(\mathrm{Var}[X_{2,i} - X_{3,i}] + \mathrm{Var}[X_{2,i} - X_{4,i}] + \mathrm{Var}[X_{3,i} - X_{4,i}]\right)\,,\end{aligned}$$

which is identical to the four-cornered hat error variance relation for $X_{1,i}$ in Eq. (5) with 0-valued error covariance terms.

Number of relations for the 3CH method

For the 3CH method with N data sets, we can write the system of variance of differences as

$$\begin{aligned}\mathrm{Var}\left[X_{1,i} - X_{2,i}\right] &= \mathrm{Var}[\varepsilon_{X1,i}] + \mathrm{Var}[\varepsilon_{X2,i}] - 2\mathrm{Cov}[\varepsilon_{X1,i}, \varepsilon_{X2,i}] \\ \mathrm{Var}\left[X_{1,i} - X_{3,i}\right] &= \mathrm{Var}[\varepsilon_{X1,i}] + \mathrm{Var}[\varepsilon_{X3,i}] - 2\mathrm{Cov}[\varepsilon_{X1,i}, \varepsilon_{X3,i}] \\ &\vdots \\ \mathrm{Var}\left[X_{1,i} - X_{N,i}\right] &= \mathrm{Var}[\varepsilon_{X1,i}] + \mathrm{Var}[\varepsilon_{XN,i}] - 2\mathrm{Cov}[\varepsilon_{X1,i}, \varepsilon_{XN,i}] \\ &\vdots \\ \mathrm{Var}\left[X_{N-1,i} - X_{N,i}\right] &= \mathrm{Var}[\varepsilon_{X(N-1),i}] + \mathrm{Var}[\varepsilon_{XN,i}] \\ &\quad - 2\mathrm{Cov}[\varepsilon_{X(N-1),i}, \varepsilon_{XN,i}]\,.\end{aligned}$$

Note that in the above, there are $N-1$ relations containing instances of $\mathrm{Var}[\varepsilon]$ for each data set X. The method of solving for error variance is to combine any two of these relations along with a third that includes the relevant two data sets that are carried along with those two relations. The total number of error variance relations $\mathcal{N}_\varepsilon$ is thus given by "$(N-1)$ choose 2," or

$$\mathcal{N}_\varepsilon = \sum_{i=1}^{N-2} i = \frac{(N-1)(N-2)}{2}\,. \tag{6}$$

Other variance relationships

The set of variances for $X_{1,i}$ can be written

$$\mathrm{Var}\left[X'_{1,i}\right] = \mathrm{Var}\left[T'_i\right] + \mathrm{Var}\left[\varepsilon_{X1,i}\right] + 2E\left[\varepsilon_{X1,i}T'_i\right]\,. \tag{7a}$$

For data sets $X'_{1,i}$ and $X'_{2,i}$, the variance of the sum is

$$\begin{aligned}\mathrm{Var}\left[X'_{1,i} + X'_{2,i}\right] &= E\left[\left(T'_i + \varepsilon_{X1,i} + T'_i + \varepsilon_{X2,i}\right)^2\right] \\ &= E\Big[4T'^2_i + \varepsilon^2_{X1,i} + \varepsilon^2_{X2,i} + 2\varepsilon_{X1,i}\varepsilon_{X2,i} \\ &\quad + 4\varepsilon_{X1,i}T'_i + 4\varepsilon_{X2,i}T'_i\Big] \\ &= 4\mathrm{Var}\left[T'_i\right] + \mathrm{Var}\left[\varepsilon_{X1,i}\right] + \mathrm{Var}\left[\varepsilon_{X2,i}\right] \\ &\quad + 2\mathrm{Cov}\left[\varepsilon_{X1,i}\varepsilon_{X2,i}\right] + 4E\left[\left(\varepsilon_{X1,i} + \varepsilon_{X2,i}\right)T'_i\right]\,.\end{aligned}\tag{9}$$

Example: 3CH with six data sets

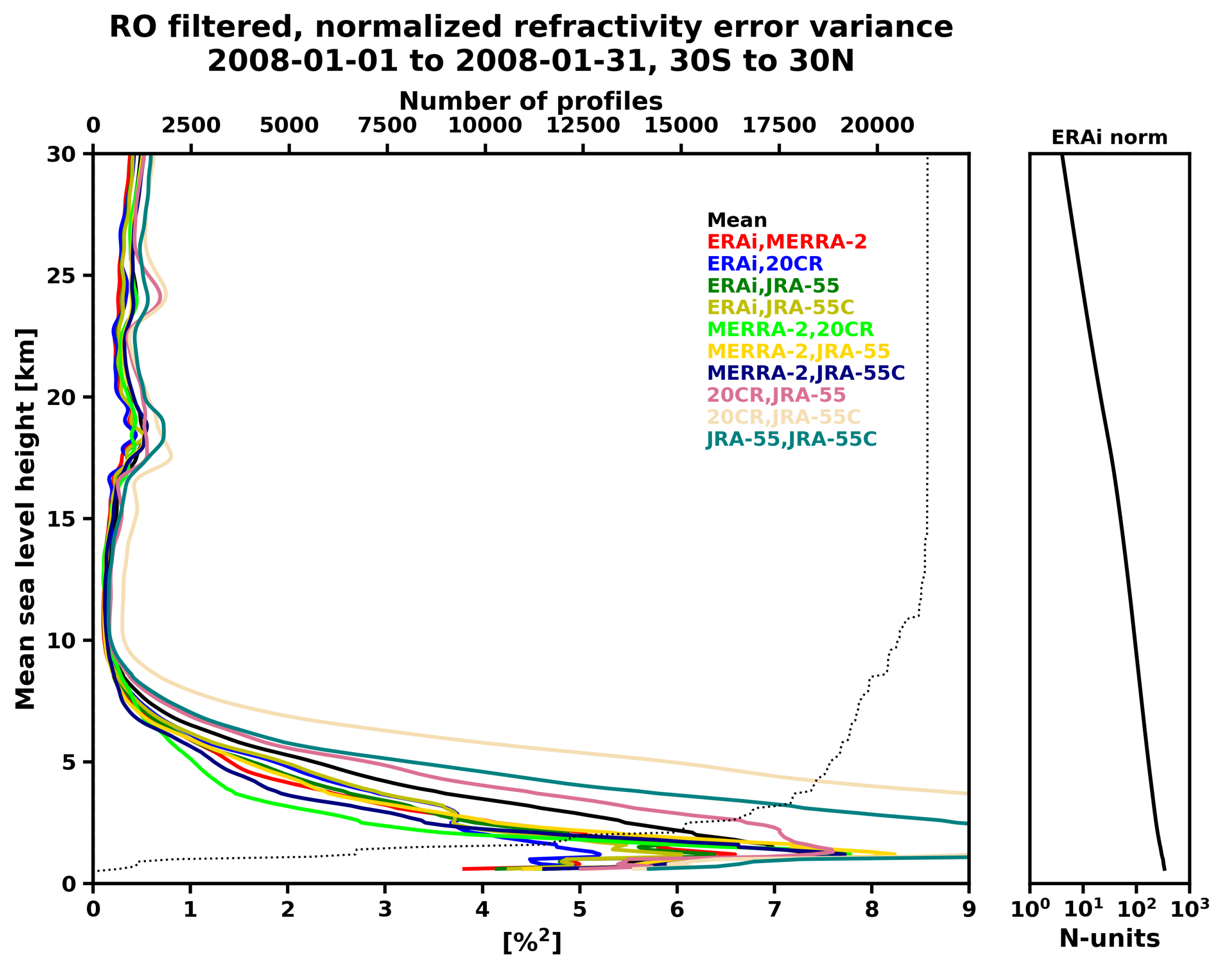

Figure 1:Normalized 3CH error variance estimates for refractivity from COSMIC radio occultation data spanning 2008-07-01 through 2008-07-31. The respective triplets are labeled by color and the mean of the ten estimates is shown in black. The number of input data points are shown by the black dotted curve. The normalization is shown on the right.

For six data sets – COSMIC RO, ERAi, MERRA-2, 20th Century Reanalysis, JRA-55, and JRA-55C – as in Figure 1, we can produce ten estimates of the error variance for a given data set. This allows us to evaluate the spread between the estimates, telling us how valid our assumption of zero error covariance is. In the above, the relatively small spread above 10 km suggests that assumption is valid while the increased spread near the surface suggests that there are non-zero error covariances.

The two-cornered hat relations

The two-cornered hat – or "triple co-location method" – error variance relation for $X_{1,i}$ is found by taking Eqs. (9)-(3a)-4·(7a):

$$\begin{aligned}&\mathrm{Var}\left[X'_{1,i} + X'_{2,i}\right] - \mathrm{Var}\left[X'_{1,i} - X'_{2,i}\right] - 4\mathrm{Var}\left[X'_{1,i}\right] \\ &\qquad = -4\mathrm{Var}\left[\varepsilon_{X1,i}\right] + 4\mathrm{Cov}\left[\varepsilon_{X1,i}\varepsilon_{X2,i}\right] - 4E\left[\varepsilon_{X1,i}T'_i\right] + 4E\left[\varepsilon_{X2,i}T'_i\right] \\ &\Longrightarrow 4\mathrm{Var}\left[\varepsilon_{X1,i}\right] \\ &\qquad = 4\mathrm{Var}\left[X'^2_{1,i} - X'_{1,i}X'_{2,i}\right] + 4\mathrm{Cov}\left[\varepsilon_{X1,i}\varepsilon_{X2,i}\right] - 4E\left[\left(\varepsilon_{X1,i} - \varepsilon_{X2,i}\right)T'_i\right] \\ &\Longrightarrow \mathrm{Var}\left[\varepsilon_{X1,i}\right] \\ &\qquad = \frac{1}{2}\left(\mathrm{Var}\left[X'_{1,i} - X'_{2,i}\right] + \mathrm{Var}\left[X'_{1,i}\right] - \mathrm{Var}\left[X'_{2,i}\right]\right) \\ &\qquad\quad + \mathrm{Cov}\left[\varepsilon_{X1,i}, \varepsilon_{X2,i}\right] - E\left[(\varepsilon_{X1,i} - \varepsilon_{X2,i})T'_i\right]\end{aligned}$$

Note that unlike the 3CH method and its generalization to N data sets, the above relies on knowledge or assumptions about the reference data set T_i.

Summary

For data sets that can be cast following Eqs. (1a-c), we may derive the 3CH relations for error variance. By assuming that there is no error covariance between the three data sets, we may apply this method to observations.

Here we show that:

- The 3CH method can be generalized for error variance estimation using N different data sets.
- The single estimate of error variance using the N data sets is equal to the mean of the $\frac{(N-1)(N-2)}{2}$ estimates of error variances for each individual data set using the 3CH method. These estimates can be used to evaluate our assumption of no error covariance between any given triplet of data sets.
- The 3CH method contains as a subset three estimates of error variance using only two sets of data – sometimes called the "triple co-location method." Application of this method requires making assumptions about the reference data set that will lead to decreased accuracy.

References

- Anthes and Rieckh (2018), DOI: 10.5194/amt-11-4239-2018
- Rieckh and Anthes (2018), DOI: 10.5194/amt-11-4309-2018