Determining the Optimum Number of Harmonics to Represent Normals Based on Multiyear Data

EDWARD S. EPSTEIN

NOAA/NWS/NMC/CAC, Camp Springs, Maryland

(Manuscript received 15 January 1991, in final form 22 April 1991)

ABSTRACT

Appropriately defined goodness-of-fit statistics are shown to provide a reasonable and objective means to determine the optimum number of harmonics to represent an annual climatology. The method is described in terms of its application, with varying degrees of success, to 5-day temperature means, their standard deviations, and to 5-day means of daily maximum and minimum temperatures.

1. Introduction

Given daily, pentad, weekly, or monthly data, the question often arises as to how much smoothing to apply to the data to produce a seasonally varying annual climatological "normal." If the smoothing is to be accomplished by the use of harmonics (and there is no guarantee that sines and cosines are optimal for describing the annual cycle), then the questions must be asked: first, how many harmonics should be used, and second, is the "best" fit an acceptable representation of the data. For example, Shea (1984) "arbitrarily" used the first five harmonics, and Liebmann et al. (1989), in a different context, questioned the propriety of more than two. Epstein and Barnston (1990) used both two and three harmonics for different variables, also "arbitrarily."

The mean datum for each calendar date (or each pentad, each month, etc.) contains both the climatological signal that one is attempting to assess, and unwanted noise of climatological, meteorological, and observational origin.[1] Underfitting (using too few harmonics) means that all of the climatological signal present in the data is not adequately included in the harmonic representation. Over much of the year the proposed climatology will prove unnecessarily biased as future data are accumulated, differing from the proposed climatology in much the same manner as in the period of record. Overfitting implies that some or much of the noise in the past data will tend not to be replicated in future observations.

A solution to this problem will be described in the context of deriving new normals of both the mean and standard deviation for 5-day mean surface temperatures at 198 stations in the contiguous United States. Data for these stations, carefully quality-controlled, all of which have less than 4%[2] of the data missing since October 1972, were made available by the Techniques Development Laboratory, NWS.

The method we have adopted works admirably for 5-day mean temperatures (section 3), less well for the standard deviation (section 4), and less well still, although still usefully, for the 5-day means of daily maximum and minimum temperatures (section 5).

2. A statistical model for the mean

Consider that Y_{ij} represents the value of the datum for the ith period (day, pentad, etc.) of the jth year, and $Y_{i.} = \sum_{j=1}^{N} Y_{ij}/N\ (i = 1, M)$ is the N-year average for each of the M periods of the year. Also let $C_i^{(k)}$ be the value of the norm as determined by the first k harmonics. In other words,

$$C_i^{(k)} = a_0 + \sum_{j=1}^{k} [a_j \cos(2\pi ij/M) + b_j \sin(2\pi ij/M)], \tag{1}$$

where

$$a_0 = (1/M) \sum_{i=1}^{M} Y_{i.},$$

$$a_j = (2/M) \sum_{i=1}^{M} Y_{i.} \cos(2\pi ij/M),$$

[1] By climatological noise we mean the unsteadiness of the climate within the period of record.

[2] Missing data were less than 1% at all but 13 of the stations used.

Corresponding author address: Dr. Edward S. Epstein, Climate Analysis Center, NWS/NOAA, 5200 Auth Road, Camp Springs, MD 20233.

and

$$b_j = (2/M) \sum_{i=1}^{M} Y_{i.} \sin(2\pi ij/M).^3$$

For each period we calculate a standard deviation

$$\sigma_i = [\sum_{j=1}^{N} (Y_{ij} - Y_{i.})^2/(N-1)]^{1/2} \qquad (2)$$

and the variable

$$T_i = (Y_{i.} - C_i^{(k)})/(\sigma_i/N^{1/2}).$$

If the Y_{ij} are drawn from populations whose true means are given by $C_i^{(k)}$ and are normally distributed, or if N is sufficiently large that the Central Limit Theorem applies, then the mean values, $Y_{i.}$, will be normal and the T_i will be distributed as Student's t with $N-1$ degrees of freedom. This is true for each i, although the T_i need not be statistically independent. If the data are underfit (consider the case where the annual mean is all that is used to represent the annual cycle of a variable that undergoes noticeable seasonal changes), then large absolute values of T_i will occur too frequently to support the hypothesis that the true means are given by $C_i^{(k)}$. On the other hand, if the data are overfit (consider the extreme case where all the means are fit exactly), there will be fewer cases of large $|T_i|$ than one would expect by chance and again the hypothesis must be rejected. Thus, the relative frequencies of the statistic T_i can be used to assess the optimum value of k for determining the annual climatology, or at least whether particular values of k are consistent with this statistical model.

The frequency of different values of T_i can be examined on a station-to-station basis or collectively. If the stations are treated individually, and there are many stations (say of the order of 100), then one must recognize the strong likelihood that some valid models will be rejected. We have adopted the procedure of initially treating all stations collectively and assuming that one value of k is satisfactory for all. This hypothesis will be rejected and an alternative proposed only if there is no value of k for which the calculated T_i fit, sufficiently closely, the proper t distribution.

3. Application to 5-day mean temperatures

For each of several values of k, and for each station, counts were made (f_m, m = 1,10) of the number of pentads for which the value of T_i fell into the mth decile of a Student's t distribution with 16 or 17 degrees of freedom, depending on the value of N for that pentad. (This assumes statistical independence of the 5-day means from one year to another.) There are 73 pentads in a year and, therefore, one expects 7.3 values in each decile. If the model is appropriate then the statistics $X = \sum_{m=1}^{10} (f_m - 7.3)^2/7.3$ (one for each station) should have a χ^2 (chi-square) distribution with nine degrees of freedom. Also the f_m were summed over all 198 stations, and the goodness-of-fit statistic X was determined for these overall frequencies.

Table 1 contains the average value of X over all stations, and also one-half the between-station variance, for values of k from 1 to 6. The expected value of a chi-square variable is equal to its degrees of freedom and its variance is twice its degrees of freedom. Therefore, if the X are indeed distributed as χ^2 with nine degrees of freedom, the entries in both columns should be nine, within the limits of sampling variability. (The sampling uncertainty of the mean can be judged by estimating the standard deviation of the mean, which is the square root of twice the number in column three divided by 198, the number of stations.) On the basis of Table 1 we can conclude that one harmonic is too few and six is too many. These statistics appear to favor the choice of either three or four harmonics, but do not exclude either two or five.

This uncertainty is resolved, however, by the frequency statistics summed over all stations, given in Table 2. It becomes clear that three harmonics is optimum. It is also evident that the use of three harmonics for the mean at all stations is reasonable. (A χ^2 value of 13.5 does not warrant rejection even at a 10% level of significance.) The use of four harmonics overfits slightly, resulting in too many small values of T_i. This tendency was too slight to be evident consistently in the histograms of the individual stations, but when the frequencies are aggregated, providing a much more powerful test, the effect becomes quite noticeable.

4. Application to the variance of 5-day mean temperatures

Thus far we have been considering the harmonic representation of the annual cycle of the 5-day mean

TABLE 1. Estimates of the degrees of freedom of the statistic X for various numbers of harmonics.

Number of harmonics	Mean value of X	Variance of X divided by 2
1	30.19	219.43
2	9.64	12.96
3	8.69	9.19
4	8.94	9.80
5	9.53	10.81
6	10.98	15.93

[3] These Fourier coefficients are numerically entirely equivalent to a least squares fit of the truncated harmonic series to the 73 data points. There are several factors, however, that make some of the theoretical statistical attributes of least squares inappropriate here. These include the lack of independence among the data, and the fact that the variance of the data about the regression surface is not constant, but varies with the season. The ability of least squares to "fit" the data and weigh most heavily the largest departures is not affected, but the statistical attributes of maximum likelihood estimates cannot be directly invoked.

TABLE 2. Aggregated frequencies of T_i falling in deciles of Student's t distribution with 17 degrees of freedom (d.f.).

	Number of harmonics					
Decile	1	2	3	4	5	6
1	2999	1718	1424	1239	1220	1090
2	1274	1434	1407	1351	1369	1299
3	1026	1206	1393	1468	1453	1383
4	973	1231	1390	1481	1500	1530
5	918	1427	1489	1516	1510	1633
6	939	1468	1475	1654	1685	1810
7	1079	1322	1458	1549	1568	1763
8	1091	1304	1467	1536	1513	1645
9	1256	1454	1533	1385	1380	1365
10	2899	1890	1418	1275	1256	936
Chi-square (9 d.f.)	4002.	284.7	13.54	106.1	125.2	507.5

temperatures. We can apply similar considerations to the harmonic representation of the standard deviation, or variance.[4] There is no obvious reason why the fit to the second moment should require more, fewer, or the same number of harmonics as are used for the mean.

Again we invoke a statistical model. First, the variance for each pentad is recalculated using Eq. (2), but replacing $Y_{i.}$, the observed mean, with $C_i^{(k)}$, the "true" mean, and $N-1$ with N. We fit these with a harmonic series as in Eq. (1). Let the hypothesis be true that the mean values are given by $C_i^{(k)}$, and further let $\sigma_i^{(k')}$ be the standard deviation determined with a fit to the variances using k' harmonics. Then, if the individual 5-day mean temperatures are normally distributed, the statistics

$$W_i = \sum_{j=1}^{N} [(Y_{ij} - C_i^{(k)})/\sigma_i^{(k')}]^2$$

will have an F distribution with N and ∞ degrees of freedom (or, equivalently, NW_i will have a chi-square distribution with N degrees of freedom).

Calculations were made, always using three harmonics to represent the mean field, but using (as k') one, two, and three harmonics to represent the variances. Values of W_i were categorized according to their appropriate decile as determined from a table of the cumulative χ^2 distribution, using the appropriate value of $N-1$ to take into account any missing data. (Because the particular tables that were immediately available omit the 0.4 and 0.6 cumulative probabilities, the third and fourth, and the fifth and sixth deciles were combined. This should have no effect on the results.) A chi-square goodness-of-fit statistic was calculated for each station and for all stations combined. The combined results, shown in Table 3, indicate clearly that two harmonics are best in terms of overall fit, but they are unsatisfactory, in absolute terms, judged by the best value of the goodness-of-fit statistic. A value greater than 40 is clearly beyond the range that would allow us to accept the validity of all aspects of our model.

A decision was made to relax the condition that the annual march of variance need be described at all stations by the same number of harmonics. Instead, whenever the value of the goodness-of-fit statistic (the equivalent of X, above) at a particular station (calculated using the nominal number of harmonics) exceeded 14.07, the χ^2 5% significance level for seven degrees of freedom, the fit was recalculated with one additional harmonic. (This seems reasonable from the perspective of any individual station, but 5% should exceed this limit by chance. Assuming there are not very large numbers of "anomalous" stations, say 10% of the total, then half or more of those for which the extra harmonic is used may be more properly treated without it.) The revised results are shown in Table 4.

TABLE 3. Aggregated frequencies of F statistics and uniform number of harmonics.

	Number of harmonics used to fit variance			
Decile	1	2	3	Nominal count
1	1908	1503	1211	1452.7
2	1372	1347	1347	1452.7
3	1293	1410	1431	1452.7
4 + 5	2566	2890	3131	2905.4
6 + 7	2480	2843	2966	2905.4
8	1258	1424	1568	1452.7
9	1313	1447	1518	1452.7
10	2337	1663	1355	1452.7
Chi-square (7 d.f.)	844.5	43.14	85.67	

[4] There is a choice to be made whether to try to fit the standard deviation or the variance. We experimented with both and found better statistical fits to the model when we were fitting the annual cycle of the variance. This may be related to the additive property of the variance.

TABLE 4. Aggregated frequencies of F statistics; one additional harmonic allowed for stations showing 5% significance at nominal harmonic level.

Decile	Nominal number of harmonics used to fit variances: 1	2	3	Nominal count
1	1593	1470	1203	1452.7
2	1311	1352	1337	1452.7
3	1321	1397	1426	1452.7
4 + 5	2763	2880	3114	2905.4
6 + 7	2706	2871	2961	2905.4
8	1337	1468	1552	1452.7
9	1479	1472	1514	1452.7
10	1944	1544	1347	1452.7
Chi-square (7 d.f.)	236.6	15.82	85.80	
Number of stations needing extra harmonic	71	18	10	

With a mix of 18 stations requiring three harmonics[5] and the remaining 180 using two harmonics, the goodness-of-fit statistic is reduced to 15.8 which, while significant at the 5% level, is not significant at the 1% level.

Considering all the assumptions made in formulating the model, plus the fact that there is no a priori reason to believe that harmonics (sines and cosines) represent the best way to describe the annual cycles of both the mean and the variance, this seems like the best possible result.

[5] Of these 18 stations, 14 are concentrated in the northwest, including 5 of 6 stations in Montana.

5. Maximum and minimum temperatures separately

A similar analysis was applied separately to 5-day means of the daily maximum and minimum temperatures, but the results were less clear. Apparently the modeling assumptions are not as valid for daily extremes as they are for the means. (This is not surprising; one should expect extremes to have nonnormal, skewed distributions.) For the minimum temperature (see Table 5), the use of three harmonics gave strong evidence of underfitting (excessive numbers of cases of the t statistic falling in the first and tenth deciles). In contrast, three harmonics seem to overfit the maximum temperatures. The use of four harmonics overfits both the minimum and maximum temperatures, and similarly the use of two harmonics underfits both. However, none of the chi-square goodness-of-fit values were in the range that supports the validity of the model as a whole. It thus appears that the optimum number of harmonics to fit the maximum temperatures is between two and three, while between three and four are optimum for minimum temperatures. The use of extra harmonics for individual significantly misfit stations gave rise to some improvement but still failed to correct for the tendencies of the bulk of the stations to be underfit. This is largely because there were so few patently misfit individual stations (7 and 5 in the case of the maximum temperatures with 2 and 3 harmonics, respectively, and 22 and 13 for minimum temperatures with 3 and 4 harmonics). With the possible exception of the minimum temperature fit with three harmonics, this implies that, from the perspective of individual stations, there is no evidence to reject the statistical model. But with a collection of many stations the goodness-of-fit test becomes so powerful that a relatively minor deficiency in the model is detectable.

TABLE 5. Aggregated frequencies of deciles of Student's t for minimum and maximum temperatures.

	Number of harmonics							
	Minimum temperatures				Maximum temperatures			
Decile	2	3	4	3*	2	3	4	2*
1	1916	1579	1321	1513	1641	1314	1169	1607
2	1389	1477	1334	1491	1309	1324	1328	1312
3	1222	1435	1507	1434	1301	1432	1524	1302
4	1252	1348	1482	1360	1408	1488	1478	1432
5	1227	1314	1446	1321	1444	1523	1611	1463
6	1296	1316	1521	1380	1481	1586	1667	1478
7	1296	1316	1521	1380	1412	1653	1618	1432
8	1407	1416	1565	1446	1334	1474	1558	1346
9	1390	1590	1443	1539	1400	1296	1271	1404
10	2059	1602	1340	1574	1724	1364	1230	1688
x^2	503.	74.5	46.1	42.6	120.	91.8	203.	94.8

* Stations whose individual values of X exceeded 16.9 (5% significance level) were recalculated with one additional harmonic.

It seems reasonable not to insist that the overall χ^2 be less than some critical value, so long as individual stations are misfit sufficiently infrequently. Instead, one can accept, as the optimum number of harmonics, the one that minimizes the overall goodness of fit.

6. Conclusions

There does not appear to be a neat, rigorous method by which to determine unambiguously an optimum number of harmonics to fit any dataset, probably because there is no reason why all seasonal climatologies should take the form of a low-order harmonic series. The method we have described should, however, be satisfactory for most purposes. Our approach relies on statistical models which, while reasonable, are nevertheless approximate. For 5-day mean temperatures, the approximations (both as regards the sinusoidal seasonal cycle and normality) apparently are very good, so the method works well. For 5-day means of maximum and minimum temperatures separately, the approximations appear to be less good and the method less satisfactory. Nevertheless, minimizing goodness-of-fit statistics calculated on the basis of such models, while not rigorous, does provide a means to guide the selection of the appropriate number of harmonics to employ, and also offers the investigator insight into the overall validity of the model.

The method is equally applicable to daily or monthly data. It is tempting to speculate that with daily (monthly) data and greater (less) resolution in time more (fewer) harmonics will be needed, but this neglects the accompanying increase (decrease) in noise. To the extent that daily data would provide no additional information other than 5-day means concerning frequencies of 3 or 4 cycles per year, analysis of daily data should yield essentially the same result as we have obtained. To the extent that monthly data filter these same frequencies (only slightly) the optimum number of harmonics might be reduced, but this should be a minor effect.

REFERENCES

Epstein, E. S., and A. G. Barnston, 1990: A precipitation climatology of 5-day periods. *J. Climate,* **3,** 218–236.

Liebmann, B., M. Chelliah and H. van den Dool, 1989: Persistence of outgoing longwave radiation anomalies in the tropics. *Mon. Wea. Rev.,* **117,** 670–679.

Shea, D. J., 1984: The annual cycle, Part I: The annual variation of surface temperature over the United States and Canada. NCAR Tech. Note TN-242, 77 pp.