Tailoring the Barnes Scheme to Your Needs (cont.)

(c) Control of detail in the analysis (via Koch *et al,* 1983)

This means that we will determine κ_0 by requiring the response function to have certain properties. First, we will specify $\gamma = 0.2$, which means we believe the data are accurate and well-distributed. (GEMPAK default value used to be 0.3). Recall that the total response of the resulting 2-pass Barnes scheme is eq. (7) from the previous lecture:

$$D_1^* = D_0 [\, 1 + D_0^{\gamma-1} - D_0^{\gamma} \,] \qquad (7)$$

where D_0 is given by eq. (2)

$$D_0 = \exp [\, - \kappa_0 (\pi/\lambda)^2 \,] \qquad (2)$$

Solving eq. (2) for λ, we have $\lambda = \pi \left[\dfrac{- \kappa_0}{\ln D_0(\lambda)} \right]^{1/2}$

and if $\lambda = 2\Delta n$, we have $2\Delta n = \pi \left[\dfrac{- \kappa_0}{\ln D_0(2\Delta n)} \right]^{1/2}$

If we take the ratio of these last two equations, $\dfrac{\lambda}{2\Delta n} = \left[\dfrac{\ln D_0(2\Delta n)}{\ln D_0(\lambda)} \right]^{1/2}$

Thus we can solve for $D_0(\lambda)$ as a function of λ and Δn only (class exercise)

$$D_0(\lambda) = [D_0(2\Delta n)]^{\left(\frac{2\Delta n}{\lambda}\right)^2} \qquad (11)$$

Now that we know $D_0(\lambda)$, we can use this in eq. (7) to determine the total response D_1^*. But we first need to determine a value for $D_0(2\Delta n)$. Koch et al determined this value by specifying that the total response D_1^* for the wavelength $2\Delta n$ should be e^{-1} ; that is, after 2 scans, the response at $\lambda = 2\Delta n$ should be ~ .37. (Note: This may be too high for many purposes.)

Working backwards via eq. (7), we can verify that this requires $D_0(2\Delta n) = 0.0064$.

We can now use eq. (2) with $\lambda = 2\Delta n$ to compute κ_0 :

$$\kappa_0 = -(2\Delta n/\pi)^2 \ln D_0(2\Delta n)$$

or

$$\kappa_0 = 5.052\,(2\Delta n/\pi)^2 \qquad (12)$$

Recall that (12) applies only for 2 passes with $\gamma = 0.2$, and $D_1^*(2\Delta n) = e^{-1}$.

Example:

Assume $\Delta n = 50$ km. Then κ_0 from (12) $= 5119$ km^2 . This compares with $(\Delta n)^2 = 2500$ km^2 . Thus using the above κ_0 value yields a smaller response for a given λ - more smoothing. [The opposite is true is you use $(2\Delta n)^2 = 10{,}000$ km^2]

Verify the following: In the first pass of this example, with $w_m = \exp[-r^2_m / \kappa]$,

we have $w_m = e^{-1}$ for $r_m \sim 71$ km, and $D_0 \leq e^{-1}$ for $\lambda \leq 225$ km.

For the 2nd pass, (with $\gamma = 0.2$), we have $w'_m = e^{-1}$ for $r_m \sim 32$ km, and $D_1^* \leq e^{-1}$ for $\lambda \leq 100$ km (which was what was specified above).

Thus we are significantly weighting only stations very close to the grid point on the 2nd pass. If this proves too noisy, could choose a larger γ or a smaller value for $D_1^*(2\Delta n)$.

(d) Choice of grid distance Δx

This choice should also be related to Δn, noting that the minimally resolved wave in the data is $2\Delta n$. However, we would like 5-6 grid points to resolve this wave and its first derivative. This suggests that

$$\Delta x < \Delta n/2.$$

In GEMPAK, Δx is recommended to be $(0.3 - 0.5)\,\Delta n$. A smaller Δx could be used, but it is not justified by the observations.

NOTE: A smaller Δx could be justified for an analysis used in NWP: If the large-scale forcing is accurate, and the physics is correct, a small-scale grid permits mesoscale (or convective-scale) features to develop in the forecast on scales much less than Δn.

(e) Quality control

[Note: this term usually means a procedure to eliminate bad or unrepresentative data, but this is not addressed; am just using the Koch *et al* sub-category.]

(i) Choice of influence or cutoff radius R_c

R_c governs the largest value of r_m that is used in eq. (1). We could use all the data without harm, since the weight w_m just goes to zero for large r_m, so R_c is used simply to save computer time. However, it has to be large enough so that the weight is nearly = 0 when $r_m = R_c$ (i.e. – the analysis is not affected).

Two ways to select R_c include:

1. Choose R_c large enough such that all grid points are affected by at least, e.g., 3 data points – so that you get a reasonable analysis in the most data sparse region. One could go further and require that R_c is large enough such that there are data in at least 3 quadrants surrounding the grid point.

2. In Koch *et al*, they select $R_c = (20\ \kappa_0)^{1/2}$. This makes $w_m = 2x10^{-9}$ when $r_m = R_c$. We can put this in terms of grid size if we assume that $\Delta x = \Delta n/2$. Then, using eq (12), we have

$$R_c = (4\Delta x)/\pi\ (20 * 5.052)^{1/2} = 12.8\ \Delta x.$$

This result yields the two ratios:

$$R_c/\Delta x = 12.8\ ; \quad R_c/\Delta n = 6.4$$

(ii) Evaluation of the analysis

A common procedure is the compute the root mean square difference (rsmd) between the observations and the analysis interpolated to the stations (i.e. – how well does the analysis fit the data?). The rmsd will usually converge toward the data (i.e., approach 0) with increasing number of passes, but we don't want it to become zero if there are errors in the data; recall this is called overfitting.

What is an appropriate value of rmsd? One guideline is that it shouldn't be less than the rms error of observation for the data you are using. Therefore, you should know this error value in order to make good decisions in your analysis procedure. Note that it is not a wise goal to get the rmsd as close to zero as possible. Why?

Figure 6 in Koch *et al* (and Fig. 3.9 in Daley) shows the effects of different values of γ on an analysis of surface data in the Great Plains. Difference fields taken between the analyses would better show how smaller scales are "built back in" as γ becomes smaller.

Miscellaneous comments on the Barnes scheme

A. The Gaussian weight function can be used for any or all 4 dimensions

Examples include:

(a) Suppose one has many x-y analyses at successive times. We can introduce "time continuity" via

$$w_{m,n} = \exp [- r^2_m / \kappa - t^2_n / \tau]$$

over n = 1(1)N time periods. We choose τ using similar procedures to the choice of κ.

(b) Use on wind profiler time-height data:

Here we use $w_{m,n} = \exp [- z^2_m / \kappa_z - t^2_n / \tau]$

(c) Could be used on 3-D x-y-z analyses; x-z cross-sections; x-y-z-t, etc.

B. Alternative strategy to determine "best" analysis procedure

In 1994, Barnes wrote a series of three papers in *J. Atmos. Oceanic Tech.* to clarify some misconceptions users had about the Barnes scheme. The primary purpose was to clarify the four "selectable parameters" κ , γ , R_c and the number of passes n for different wavelengths and sampling distributions.

Barnes considered "well-sampled waves" to have 9 observations per wavelength if only that original field was desired, but if derivatives were calculated from the analysis, 13-15 observations per wavelength were required to prevent them from being too noisy.

Barnes defined "marginal under-sampling" as 5 obs. per wavelength, and true under-sampling to be ≤ 4 observations. It is possible to produce an acceptable analysis under these conditions if κ is chosen to be large, and you can accept noisy derivatives.

He shows how different station distributions affect the analysis. In general, a purely random distribution of stations will have 100-300% more error than the same analysis procedure applied to an evenly-spaced network.

Barnes' strategy to produce the "best analysis" is:

1. Choose a domain that is as regularly sampled as possible. Have observations outside the domain. Create synthetic or "bogus" observations in data sparse

regions to avoid uneven data distributions. (Could create these from a first pass with very large R_c and κ that produces a smooth field.)

2. Compute the uniformity ratio = $(\Delta n_r - \Delta n_c)/\Delta n_c$, where Δn_c and Δn_r are defined on p. 6 in Lecture #10.

   The uniformity ratio is ≥ 0 since $\Delta n_r \geq \Delta n_c$. [Barnes used mean distance to 6 nearest stations to compute Δn_c]

   The station distribution is "quasi-random" if the uniformity ratio is ~ 1.1.

   Note: Obs. that are *very* close together will adversely affect the Δn_c calculation. Often such obs. are averaged together to create a "super-ob".

3. Determine the "bandwidth" one has in your data domain. For example, if you have only 5-6 stations across the domain width, you essentially have no bandwidth, since the shortest possible wavelength for which you can get derivative information is also the longest. Recall that the longest or "fundamental wave" is the domain dimension; the shortest resolvable wave is $2\Delta n_r$, with $\lambda \geq 2\Delta n_r$ needed to get derivative information.

4. Perform tests with an analytic function that contains the permissible waves. Use the observed station distribution and choose a large R_c. Explore a reasonable range of κ, γ, and n values.

5. Compute the mean absolute error at grid points (m.a.e.g) for the different analyses and their derivatives. We can do this rather than the rmsd at stations because we know the truth at the grid points via the analytic fields. Thus it is permissible to get the m.a.e.g as small as possible. Determine what combination of κ, γ, n values minimizes the m.a.e.g.

Barnes (and others) have found that 3-4 passes with $\gamma = 1$ (no accelerated convergence) yields the best analysis for the purpose of computing derivatives. This leaves only κ to determine, and it is selected to minimize m.a.e.g.

Fig. 3.9 in Daley

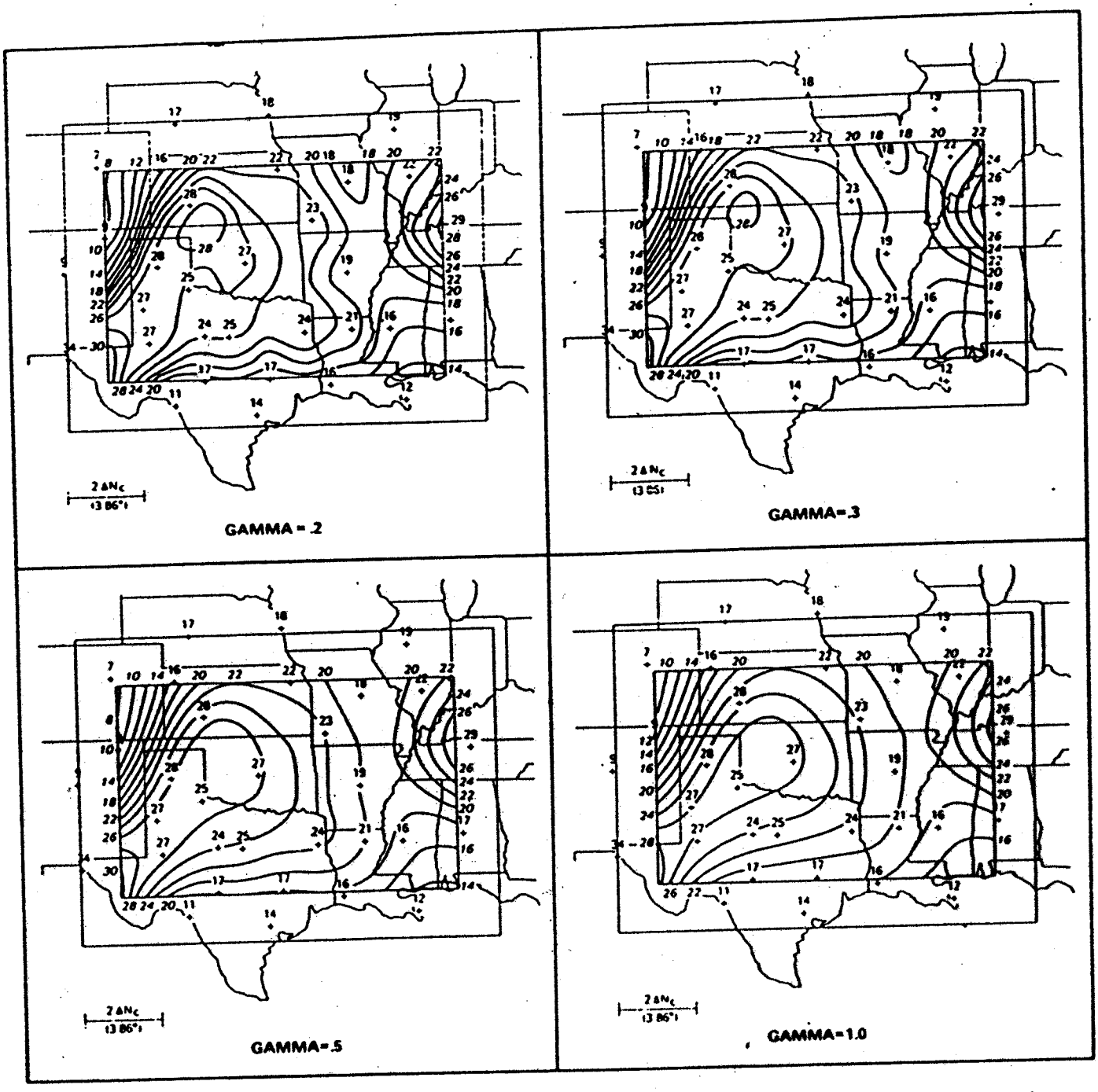

FIG. 6. Changes in the objective analysis of the SESAME RAWIN data set brought about by variations in the value input for γ (analysis with $\gamma = 0.3$ is identical to that in Fig. 5a).