40 YEARS OF DATA ASSIMILATION AT NMC/NCEP—FROM HOUGH ANALYSIS TO GRIDPOINT STATISTICAL INTERPOLATION (GSI)

David F. Parrish
Mesoscale Modeling Branch
Environmental Modeling Center
National Centers For Weather and Climate Prediction

outline of talk

- Hough function Flattery analysis
- Machenhaur Baer NLNMI
- NLNMI
- OI Bergmann, Kistler
- INLNMI
- A Lorenc / N. Phillips (formalization of DA and NGM)
- S. Cohn Kalman filter
- N. Phillips (1-d spectral model to understand OI)
- Spectral OI
- SSI
- recursive filter/radar data eta analysis
- excursion to EnKF with J. Anderson
- GSI
- Strong constraint
- 3denvar
- optimization
- multigrid for correlation/localization

1975—Flattery Hough function analysis

Prior to starting work at the National Meteorological Center (NMC, now NCEP), I worked in the Development Branch of the Air Force Global Weather Center (AFGWC, now AFWA). My boss was Col. Tom Flattery. He had recently returned from being an Air Force liaison visiting scientist at NMC. While there, he constructed a data analysis program, the Hough analysis, which performed a weighted least squares fit of observations of height and wind to Hough functions and vertical eigenfunctions resolved at a rhomboidal truncation at zonal wave number 24.

1975—Flattery Hough function analysis

This was running in operations at NMC, but there were many problems, and no one available with any knowledge about the Hough analysis code besides Col. Flattery. I was in early stages of developing a global spectral model for use at AFGWC, but would leave soon, looking for a post doctoral position. Col. Flattery shared with me the problems with the NMC Hough Analysis and decided that because of my background with spectral modeling work, I might be able to help.

1975—Flattery Hough function analysis

So one thing led to another and by February, 1975 I was working in the NMC Automation Division. My initial assignment was to fix Flattery's Hough Analysis. This turned out to be rather difficult, not because of the math, but the code, which was very difficult to decipher. So I ended up writing a complete new code which generated the same results, and then I was able to figure out what some of the problems were.

1975—Flattery Hough function analysis

The Hough functions are eigenfunctions of shallow water equations on a rotating sphere and can be divided into slowly evolving balanced modes and fast gravity modes. The slow modes have frequencies which go to zero as the rotation rate decreases to 0, while the gravity modes maintain time dependence. The Flattery analysis used only the slow modes in the analysis.

1975—Flattery Hough function analysis

Flattery used 24 east-west wave numbers and 24 latitude dependent slow modes for each east-west wave number. Height and wind were defined at 12 isobaric levels and relative humidity at the 6 lowest levels. The vertical treatment consisted of computing EOF functions for each analysis.

1975—Flattery Hough function analysis

I don’t remember many of the details, but one in particular appeared to be the main cause of problems. There was very little data in the Southern Hemisphere, and only a single radiosonde at the south pole. The analysis was computed as an incremental correction to the forecast, similar to today, but the Hough functions enforced a strict linear balance between height and wind in areas away from observations.

1975—Flattery Hough function analysis

If the south pole observation was unavailable for a few cycles (not uncommon), the analysis increments were still non zero, influenced by observations far away (because of the low resolution), and the linear balance caused a growing bias in the forecast over the south pole , so that when the south pole observation became available, it would fail the gross error check. This was fixed by overriding the south pole gross error check.

1977 Machenhauer, Baer
Nonlinear normal mode initialization (NLNMI)

This was a very exciting time, when all everyone could talk about was this new idea for creating a balanced initial state, and is also related to Hough functions, in the sense that initial model time tendencies are projected onto the fast gravity wave modes, and then a simple correction is made in mode space to make the gravity mode tendencies zero. This correction is transferred back to physical space, and added to the original state.

1977 Machenhauer, Baer
Nonlinear normal mode initialization (NLNMI)

Because the model is nonlinear, new tendencies are computed, and a new correction is found in mode space. Typically, satisfactory convergence is achieved in 2-3 iterations for deep fast modes, which are responsible for most of the initial imbalance after assimilating observations. One day in the winter of 1979, I was so excited about a new idea related to NLNMI that I drove to work very early in a blizzard and got snowed in at the office, which was OK, because the computers were still running just fine.

1981 Optimum Interpolation (Bergmann, Kistler)

The Hough analysis was replaced with an optimum interpolation (OI) analysis, and also, around this time, a global spectral model was implemented, using NLNMI.

1983 Implicit Nonlinear Normal Mode Initialization (INLNMI)

The initial forms of NLNMI depended on some transform from grid to spectral space (spherical harmonics over full global domain, double sine series over regional domain). Around this time, I figured out how to do initialization over a limited area without the requirement for expansion in horizontal modes (there is always a vertical mode transform involved).

1983 Implicit Nonlinear Normal Mode Initialization (INLNMI)

Then, independently, starting in 1985 or so, while on a visit to the Canadian met center, Clive Temperton came up with a similar idea, but with much better and more complete mathematical definitions. The code he left behind became a valuable asset again many years later (with the Canadian independent application of TLNMC, the tangent linear normal mode constraint).

1985 Andrew Lorenc visit

In 1985, Andrew Lorenc visited NMC for a year. During his visit, he did considerable work developing a standard framework for data assimilation, which unified successive correction methods, optimum interpolation, Kreiging, and various other data fitting methods.

During this time, I was creating an INMI for the NGM model, a regional model created by Norm Phillips, principle scientist at NMC.

1986-87 Steve Cohn collaboration—Kalman filter

My supervisor at this time, Ron McPherson, heard about something called the Kalman filter, and would I take some time to find out about it and if it was something we should be considering for future implementation. So I began a great collaboration with Steve Cohn, who was a post doc at NYU. About once a month I would take the train to NYC and we would spend time developing his ideas on application of Kalman filtering to atmospheric data assimilation.

1986-87 Steve Cohn collaboration—Kalman filter

For his PhD, he used a 1-d shallow water model, and now wanted to extend it to a 2-d version and experiment with ways to make practical approximations. So I spent most of my time back at NMC creating and running 2-d Kalman filtering experiments. We eventually published these results. Later, Ricardo Todling, then working on his PhD, took my program and extended it in the vertical to 2 levels. (Ricardo has had a huge influence in recent years with the evolution of GSI—more on that later).

1987-88 Norm Phillips investigates OI with 1-d model

Norm Phillips didn't have a very good idea about OI, so he created his own 1-d periodic system and looked at it from the point of view of spectral transform space. He made an assumption about model error, being distributed evenly across all wave numbers.

1987-88 Norm Phillips investigates OI with 1-d model

He found that after transformation to physical space, he obtained qualitatively similar correlation functions (in particular cross correlations between height and wind) compared to those obtained operationally from curve fitting to o-g lagged correlations with assumption of isotropy. When he was explaining this to me one day in his office, I suddenly jumped up and told him this could be applied directly to global analysis. We could do OI using a spectral representation, using the Hough functions once again as the basis.

1987-88 Spectral OI

I immediately dropped everything else I was working on and began to create this new system, which I called Spectral Optimum Interpolation (Spectral OI). For the next year, I tried many times to explain what this was, and how it was in principle equivalent to the existing OI being used at this time by most of the NWP centers.

1987-88 Spectral OI

But no one seemed to understand what I was talking about, until once at a workshop in Virginia, Oliver Talagrand suddenly jumped up in the middle of my presentation and got very excited. He was the first one to understand what I was doing (after Norm Phillips of course, who gave me the idea in the first place).

1989-92 SSI (Spectral Statistical Interpolation)

I was invited up to Princeton to give another talk about Spectral OI by John Derber, who was then a post doc at GFDL. He understood immediately what I had created. We had quite a lively discussion in his office about the possibilities and how to extend this idea into an operational system for global models. Unfortunately, my seminar was much less exciting (I am not particularly great either at writing up or presenting results of my work).

1989-92 SSI (Spectral Statistical Interpolation)

Not long after that, John accepted a position at NMC, and we began to work together in earnest to develop a new DA system for the operational global spectral model. First he suggested we change the name from Spectral Optimum Interpolation (SOI) to Spectral Statistical Interpolation (SSI), because SOI was already in use by the climate people for Southern Oscillation Index.

1989-92 SSI (Spectral Statistical Interpolation)

The next change we agreed on was that it was more practical to use spherical harmonics in place of Hough functions, and formulate coupling relationships between different variables.

1989-92 SSI (Spectral Statistical Interpolation)

We eventually were able to put together a demonstration DA system with the operational global spectral model and the new SSI analysis. The first thing that we found was that for a single analysis with identical guess and observations, the SSI based analysis increment was much smoother than the OI analysis increment. It also gave a substantially better overall fit to the observations and used less computer time.

It was put into a parallel test and was implemented operationally in 1991.

1989-92 SSI (Spectral Statistical Interpolation)

A while after implementation, I was in the hospital for a week with pancreatitis (had my gall bladder removed a few weeks later). Eugenia Kalnay, our new director, visited me in the hospital and said we were beating the European center in anomaly correlation scores. She said it was all because of me, but I pointed out to her that without John Derber (his ideas, and his skills at getting from a toy system to operations), this never would have become operational. So he definitely gets as much if not more credit. And there were lots of other people involved also—a team effort.

1989-92 SSI (Spectral Statistical Interpolation)

Another interesting thing happened after implementation of SSI. People from the ECMWF started visiting a lot, wanting to know all the details of our new scheme. Around 1995 or so, John Derber spent a year at ECMWF, helping to install something like SSI into their DA system.

He was supposed to fix it so it would always be a little worse than our system, but I guess that didn't work out (NOT TRUE—just us poking fun at John). And of course the EC moved back into the lead with best model in the world.

recursive filter/radar data eta analysis

In 1995, Geoff DiMego was hoping that we could have an analysis and forecast for the 1996 Olympic Games in Atlanta, GA. There existed an operational 88D radar, back when the current radar network was first being brought online. Geoff was pushing for us to be the first to use this data in a local analysis and forecast.

recursive filter/radar data eta analysis

I was already working with Jim Purser on a regional analysis code that was based on the SSI, but instead of representing the background error in spectral space, we used recursive filters then being developed by Jim. The eta model (Mesinger and Janjic) was being used now as the operational regional model.

recursive filter/radar data eta analysis

The first radar data we used was not directly from the radars, but first processed into a highly compressed format (NIDS) that was used to make local and national graphic displays of reflectivity and radial wind. I was able to initially download this information through a receiver and computer box near my office. I had to do some very careful programming on the computer attached to the receiver and transfer it off quickly enough to keep from loosing the data.

recursive filter/radar data eta analysis

This was useful for the early development of using radar data in the regional model analysis. I was also able to get a direct connection to the radar at Atlanta using ftp.

Although some special forecasts were made for the Atlanta games, we were not able to get the radar data in to the system in time.

anisotropic recursive filter

By this time, Jim Purser had constructed an elaborate mathematical structure for using recursive filters to generate anisotropic correlation functions. Riishojgaard had demonstrated a simple technique that could be used with traditional OI where correlations could be stretched along contours of some field, such as temperature gradients along frontal boundaries.

anisotropic recursive filter

Using Purser's more efficient recursive filters, adapted by me for anisotropic use in the eta 3dvar, Manuel Pondeca did a lot of detailed experiments.

Unfortunately, these have not been used operationally, except in the Real Time Mesoscale Analysis (RTMA), which obtains hourly analyses of various surface fields (psfc, T, q, wind, wind gusts, ceiling and visibility, etc.) More on this later in the talk.

excursion to EnKF with J. Anderson

In 2001 I got very excited at a seminar by Jeff Anderson, about his ideas on how to use the Ensemble Kalman Filter. He had a very simple (high school math) explanation that he implemented in that form (which was equivalent to a sqrt Kalman filter). He showed very interesting results for a full multilevel global model with physics, in which he assimilated only surface pressure. The results were quite impressive, but was for an idealized case, not real data.

excursion to EnKF with J. Anderson

I decided I had to give this a try as a little side project for my own education. So I got Mark Iredell to provide me with scripts to run the GDAS (global data assimilation system), but replaced the analysis with the Anderson version of EnKF. I ran experiments with real data, first at T64 and later T96 resolution.

excursion to EnKF with J. Anderson

The results were interesting, but a bit disappointing. I had some long conversations with Jeff on a visit to NCAR where we decided that his ideal experiments were not replicated with real data, probably because of unknown model bias relative to the real atmosphere.

GSI (Gridpoint Statistical Interpolation)

While I was wasting time with the Anderson version of EnKF, Wan-Shu Wu was working in her systematic and careful way on the new GSI code. Around 1998 (?) John Derber set up a morning meeting at his home where John, Wan-Shu, and myself worked out details of what a grid point version of SSI, using Jim Purser's recursive filter ideas, would look like.

GSI (Gridpoint Statistical Interpolation)

In 2002, we had a DA workshop, and then a smaller meeting to decide on the next phase of the experimental GDAS system, based on Wan-Shu's new grid point formulation, the GSI, based on Jim Purser's recursive filters. At that meeting, I made a proposal that the new GSI not be limited to just the global model, but that it also work for regional models. The reason for this was that I was the primary person responsible for the regional eta model analysis, which was different in many details, and which contained my adaptation of the SSI satellite radiance data part of the code.

GSI (Gridpoint Statistical Interpolation)

My attempt to keep up with global radiance processing was always about 2 years behind. Everyone agreed that this was an excellent idea. This was the seed that led to a unified code that has now evolved into a large collaborative effort.

Tangent Linear Normal Mode Constraint (TLNMC, aka Strong Constraint)

In the summer of 2006, I was on vacation at a lake in Georgia with lots of relatives. As is often the case, I kind of ruined the vacation for everybody because I had just had a really big idea for improving balance in the GSI, which was then not yet operational. I spent most of the time locked in a room by myself making derivations for what became a novel method for applying nonlinear normal mode initialization to the analysis increment during the analysis iterations, instead of to the full analysis fields output from GSI.

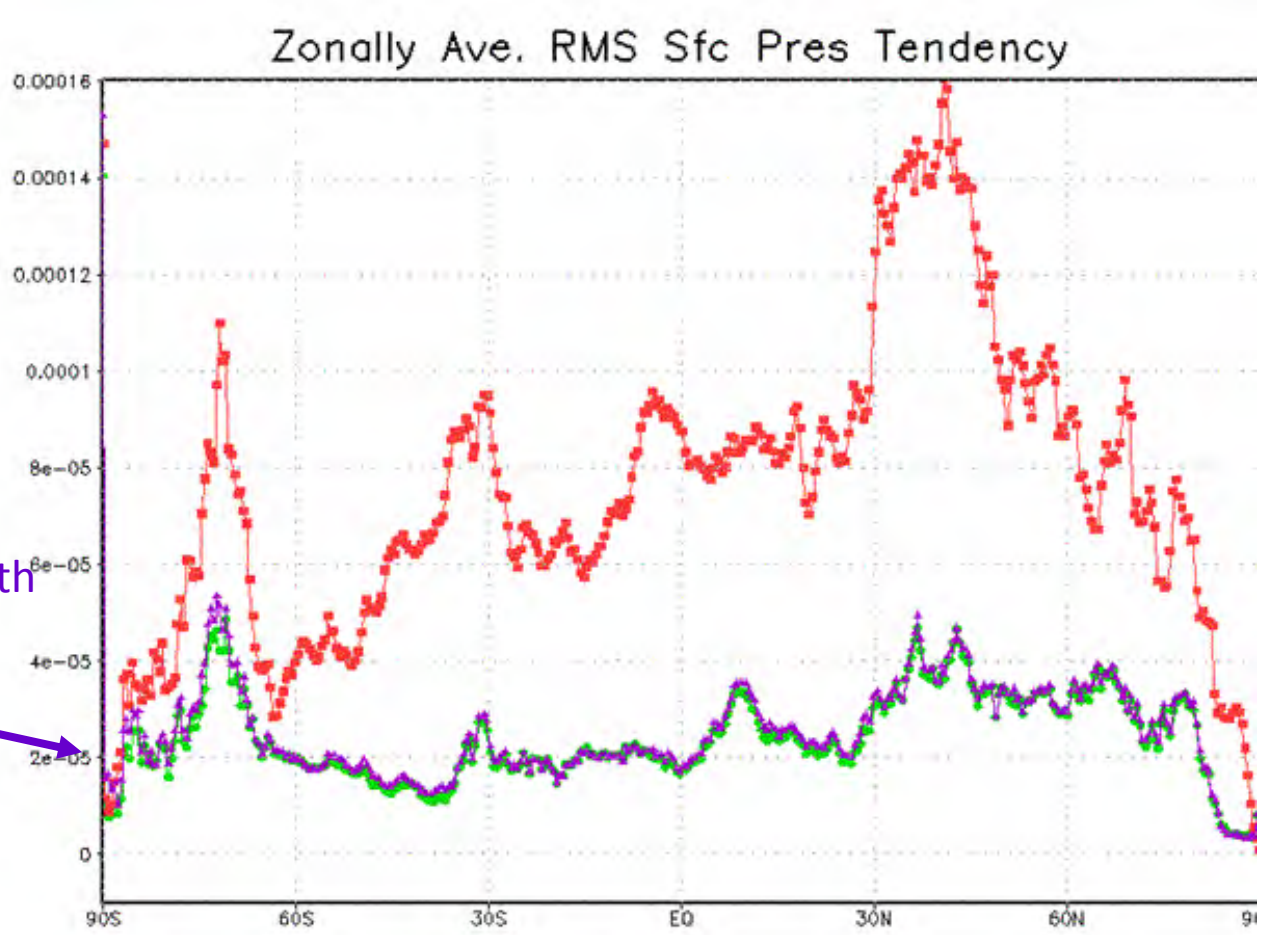
Tangent Linear Normal Mode Constraint (TLNMC, aka Strong Constraint)

When I got back from vacation, just like back in 1988, I ran around to everyone trying to explain this idea and how it would work. The main thing was that we needed tangent linear and adjoints of the model. Daryl Kleist and I decided to give this a try. Daryl was already working on a generic global spectral model (dynamics only) tangent linear and adjoint code for a weak constraint he was adding to GSI.

Tangent Linear Normal Mode Constraint (TLNMC, aka Strong Constraint)

I put together a version of normal mode initialization that would project the analysis increment at every iteration as the first part of the forward model from analysis variable to observation. This way the observation residual would see only a balanced increment to compare against.

Surface Pressure Tendency Revisited



Minimal increase with TLNMC

Zonal-average surface pressure tendency for background (green), unconstrained GSI analysis (red), and GSI analysis with TLNMC (purple)

Tangent Linear Normal Mode Constraint (TLNMC, aka Strong Constraint)

There ended up being several of us involved in this project, which helped Daryl get a well deserved PhD. Daryl also wrote a paper, with lots of co-authors. It is interesting that the very same idea was developed independently at about the same time by Luc Fillion and others for the Canadian regional model.

Tangent Linear Normal Mode Constraint (TLNMC, aka Strong Constraint)

Theirs was much easier to do, since they already had tangent linear and adjoint models and the great normal mode machinery installed by Clive Temperton during an extended visit many years earlier.

Tangent Linear Normal Mode Constraint (TLNMC, aka Strong Constraint)

As it happened, this was a necessary component added to GSI in order to obtain improved results over the operational SSI system, a requirement for implementation. The unfortunate part about this is that we now need the TLNMC, even though it probably causes some degradation in the tropics, due to the absence of moist physics.

Tangent Linear Normal Mode Constraint (TLNMC, aka Strong Constraint)

Several people in the GSI development group (Cathy Thomas, Emily Liu, Yanqiu Zhu and others) worked very hard to add moist physics from the operational spectral model, but apparently parallel tests yielded essentially neutral results with no improvement in the tropics, and the added code made the already expensive TLNMC much more costly. This was a rather depressing and demoralizing result.

3denvar

Starting in 2010 I decided, in my usual impulsive manner, to go ahead and install a version of hybrid ensemble 3dvar as an added option in GSI. It appeared that this was an important development and we should get a jump on this idea. This was a good idea in retrospect, since we were able to precede other centers with operational implementation of 3denvar.

3denvar

There was a convergence problem that appeared in the regional 3denvar, which uses the global 3denvar passively (Wan-Shu installed this and through her careful attention to detail was able to obtain significant improvement in the regional system). Convergence warnings appeared for analyses over small domains with only a few observations. This only occurred when the option was turned on to allow vertically varying beta weights between static background and ensemble error covariances. I quickly realized that this was because the beta weights were originally assumed to be a scalar constant multiplying B and A.

3denvar

However, adding vertical variation means that now beta1, beta2 are diagonal matrices which do not commute with B or A. So now the covariance matrices are no longer symmetric, as required for the minimization algorithm. To fix this was straightforward, although I managed to make several mistakes and took longer than usual to get this fix in.

This fix is now in line to get into the GSI trunk.

3denvar

There are other, independent, convergence problems in the PGSOI minimization option that John Derber has been working on.

Optimization of GSI

When Wan-Shu, Daryl and I were here 3 years ago, I presented information on a communication code I was developing to make GSI more scalable to larger number of processors with increasing resolution. GSI was originally designed with 2 storage modes: (1) rectangular subdomains in the horizontal with all vertical levels/variables available on each processor, and (2) full horizontal domains with vertical variables/levels distributed across processors.

Optimization of GSI

This worked well when there were a small number of nodes, not very many processors per node, and very large memory on a node. Now, storage of full horizontal fields has resulted in an upper limit to resolution for GSI. The upcoming implementation of the T??? 4denvar GSI uses 240 nodes with only 2 processors/12 threads per node on our new Cray computer and takes 25 minutes to run.

Optimization of GSI

This module, genex_mod.f90, was designed as a replacement for general_sub2grid_mod.f90. However, many interruptions, related to correcting various problems with GSI, has limited work on this optimization project.

Optimization of GSI

So far there are only two subroutines, psichi2uv_regional.f90 and its adjoint, which have been successfully converted to use only subdomains with single row halo updating by genex routines. Unfortunately, there is almost no savings from these modifications because the fraction of total run time is very small already.

Optimization of GSI

A very large improvement in performance could come from modifications to the horizontal correlation and localization routines. I tried this initially for the global version, but it was not practical due to interpolation between lat-lon and polar stereographic caps for application of recursive filters.

Optimization of GSI

And even for regional analyses, some improvement may be obtained by switching between subdomain storage and all x-direction followed by all y-direction then back again with adjoint. For the anisotropic recursive filter used in the 2D rtma, it already runs in subdomain mode, but is completely communication bound due to gathering of points from subdomains to groups of strings of points running obliquely with variable lengths.

Optimization of GSI

An alternative that I have explored with help from Wan-Shu is to use a multigrid method to efficiently compute the direct multiplication of a correlation matrix with a vector. I was able to create an anisotropic demonstration program in 2D for 1 processor that was applied to the RTMA.

Optimization of GSI

This past year, I wanted to create a subdomain version of the covariance/localization code, but have only made a little bit of progress so far because of other more immediate issues with GSI. These were, first the correction to fix convergence errors in 3d and 4denvar, previously mentioned, and now an unexpected application of the genex module to speed up reading of ensemble perturbations. This last could be the first application of genex that will save from 3-5 minutes from the run time for the 4denvar scheduled for implementation in the 3rd quarter of this fiscal year.

Hybrid EnVar (fix for variable β_f & β_e)

Lorenc (2003), Buehner (2005), Wang et al.(2007)

(thanks to Daryl Kleist for letting me borrow this from his presentation)

$$J_{Hyb}(x_c,\alpha) = \beta_c \frac{1}{2}(x_f')^T B(x_f') + \beta_e \frac{1}{2}(\alpha)^T L^{-1}(\alpha) + \frac{1}{2}(y_o' - Hx_t')^T R^{-1}(y_o' - Hx_t')$$

$$\mathbf{x}'_t = \mathbf{x}'_f + \sum_{m=1}^{M}(\alpha^m \circ \mathbf{x}_e^m)$$

1. β_f & β_e are now allowed to vary in the vertical. This creates a problem with the existing GSI code, because the one sided multiply makes the effective matrix non-symmetric. To fix this, we multiply on the right and left of the symmetric matrices **B** and **L** by 1/sqrt(β_f) and 1/sqrt(β_e) respectively.

For multiple localizations at different scales, can add multiple copies of ensemble control variables with desired band pass properties.