NASA/TM–2018-104606 / Vol. 50

## Technical Report Series on Global Modeling and Data Assimilation, Volume 50

*Randal D. Koster, Editor*

# The GMAO Hybrid Ensemble-Variational Atmospheric Data Assimilation System: Version 2.0

*Ricardo Todling and Amal El Akkraoui*

National Aeronautics and
Space Administration

**Goddard Space Flight Center**
**Greenbelt, Maryland 20771**

**March 2018**

# NASA STI Program ... in Profile

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA scientific and technical information (STI) program plays a key part in helping NASA maintain this important role.

The NASA STI program operates under the auspices of the Agency Chief Information Officer. It collects, organizes, provides for archiving, and disseminates NASA's STI. The NASA STI program provides access to the NASA Aeronautics and Space Database and its public interface, the NASA Technical Report Server, thus providing one of the largest collections of aeronautical and space science STI in the world. Results are published in both non-NASA channels and by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA Programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counterpart of peer-reviewed formal professional papers but has less stringent limitations on manuscript length and extent of graphic presentations.
- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.
- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.
- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.
- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.
- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services also include organizing and publishing research results, distributing specialized research announcements and feeds, providing help desk and personal search support, and enabling data exchange services. For more information about the NASA STI program, see the following:

- Access the NASA STI program home page at http://www.sti.nasa.gov
- E-mail your question via the Internet to help@sti.nasa.gov
- Phone the NASA STI Information Desk at 757-864-9658
- Write to:
  NASA STI Information Desk
  Mail Stop 148
  NASA's Langley Research Center
  Hampton, VA 23681-2199

NASA/TM–2018-104606 / Vol. 50

**Technical Report Series on Global Modeling and Data Assimilation, Volume 50**

*Randal D. Koster, Editor*

# The GMAO Hybrid Ensemble-Variational Atmospheric Data Assimilation System: Version 2.0

*Ricardo Todling*
*NASA's Goddard Space Flight Center, Greenbelt, MD*
*Amal El Akkraoui*
*Science Systems and Applications, Inc., Lanham, MD*

National Aeronautics and
Space Administration

**Goddard Space Flight Center**
**Greenbelt, Maryland 20771**

**March 2018**

*Level of Review: This material has been technically reviewed by technical management.*

Available from

NASA STI Program
Mail Stop 148
NASA's Langley Research Center
Hampton, VA 23681-2199

National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161
703-605-6000

# Abstract

This document describes the implementation and usage of the Goddard Earth Observing System (GEOS) Hybrid Ensemble-Variational Atmospheric Data Assimilation System (Hybrid EVADAS). Its aim is to provide comprehensive guidance to users of GEOS ADAS interested in experimenting with its hybrid functionalities. The document is also aimed at providing a short summary of the state-of-science in this release of the hybrid system. As explained here, the ensemble data assimilation system (EnADAS) mechanism added to GEOS ADAS to enable hybrid data assimilation applications has been introduced to the pre-existing machinery of GEOS in the most non-intrusive possible way. Only very minor changes have been made to the original scripts controlling GEOS ADAS with the objective of facilitating its usage by both researchers and the GMAO's near-real-time Forward Processing applications.

In a hybrid scenario two data assimilation systems run concurrently in a two-way feedback mode such that: the ensemble provides background ensemble perturbations required by the ADAS deterministic (typically high resolution) hybrid analysis; and the deterministic ADAS provides analysis information for recentering of the EnADAS analyses and information necessary to ensure that observation bias correction procedures are consistent between both the deterministic ADAS and the EnADAS. The nonintrusive approach to introducing hybrid capability to GEOS ADAS means, in particular, that previously existing features continue to be available. Thus, not only is this upgraded version of GEOS ADAS capable of supporting new applications such as Hybrid 3D-Var, 3D-EnVar, 4D-EnVar and Hybrid 4D-EnVar, it remains possible to use GEOS ADAS in its traditional 3D-Var mode which has been used in both MERRA and MERRA-2. Furthermore, as described in this document, GEOS ADAS also supports a configuration for exercising a purely ensemble-based assimilation strategy which can be fully decoupled from its variational component.

We should point out that Release 1.0 of this document was made available to GMAO in mid-2013, when we introduced Hybrid 3D-Var capability to GEOS ADAS. This initial version of the documentation included a considerably different state-of-science introductory section but many of the same detailed description of the mechanisms of GEOS EnADAS. We are glad to report that a few of the desirable Future Works listed in Release 1.0 have now been added to the present version of GEOS EnADAS. These include the ability to exercise an Ensemble Prediction System that uses the ensemble analyses of GEOS EnADAS and (a very early, but functional version of) a tool to support Ensemble Forecast Sensitivity and Observation Impact applications.

# Contents

# List of Tables

# List of Figures

# 1 Hybrid-Ensemble Data Assimilation

## 1.1 Introduction

The basic idea of hybrid variational data assimilation is to use an ensemble of background fields to introduce instantaneous, flow-dependent, features to the traditionally non-evolving (climatological) background error covariance of three-dimensional variational (3D-Var) systems. Generally, following the notation in Lorenc *et al.* (2015), the cost function of a four-dimensional (4D) analysis procedure can be written as

$$J(\delta\underline{\mathbf{x}}) = \frac{1}{2}\delta\underline{\mathbf{x}}^T\underline{\mathbf{B}}\delta\underline{\mathbf{x}} + \frac{1}{2}(\underline{\mathbf{y}}-\underline{\mathbf{y}}^o)^T\underline{\mathbf{R}}^{-1}(\underline{\mathbf{y}}-\underline{\mathbf{y}}^o) + J_c\,, \tag{1}$$

where the first term is a regularization term that takes into account prior, background, information, the second term relates to the guess fit to the observations, and the last term is associated with possible additional constraints imposed on the analysis, such as an imbalance penalty term like that of Gauthier and Thépaut (2001) or dry mass conservation term (e.g., Takacs *et al.* 2016). The underbars in the vector and matrix in the expression above amount to a compact notation folding the time dimension into a single vector quantity. For example, the incremental solution vector $\delta\underline{\mathbf{x}} = (\delta\mathbf{x}_0^T, \delta\mathbf{x}_1^T, \cdots, \delta\mathbf{x}_K^T)^T$ is obtained from a minimization problem covering the time interval $[t_0, t_K]$ that includes $K+1$ discrete time slots.

Thus, the minimization of the cost function (1) seeks to find a 4D incremental correction $\delta\underline{\mathbf{x}}$ to a 4D background state $\underline{\mathbf{x}}^b$ using observations $\underline{\mathbf{y}}^o$ covering the time interval associated with the 4D problem. The solution depends on the fit of observations to the model-predicted observations $\underline{\mathbf{y}}$ obtained from a nonlinear observation function $\underline{H}$ as in

$$\underline{\mathbf{y}} = \underline{H}(\underline{\mathbf{x}}^b + \delta\underline{\mathbf{x}})\,, \tag{2}$$

and on the 4D weighing error covariance matrices $\underline{\mathbf{B}}$ and $\underline{\mathbf{R}}$ associated with the background and observations terms in the cost function, respectively.

The present documentation describes the implementation of the so-called Hybrid four-dimensional ensemble-variational (4D-EnVar) approach in the Goddard Earth Observing System (GEOS). In this, the incremental solution to the minimization of $J$ is a linear combination of the solution $\delta\mathbf{x}$ of the standard 3D-Var problem with a 4D component that comes from an $M$-member ensemble. That is,

$$\delta\underline{\mathbf{x}} = \beta_c\delta\mathbf{x} + \beta_e\sum_{m=1}^{M}\underline{\alpha}_m \circ \delta\underline{\mathbf{x}}_m^e\,. \tag{3}$$

Here, the symbol $\circ$ stands for the Hadamard-Schur (element-wise) product of two vectors, $\underline{\alpha}_m$ is the $m$-th 4D control vector related to the $m$-th 4D ensemble member[1], and $\delta\underline{\mathbf{x}}_m^e = (\underline{\mathbf{x}}_m - \bar{\underline{\mathbf{x}}})/\sqrt{M-1}$ is the $m$-th ensemble perturbation created from the $m$-th member 4D background state $\underline{\mathbf{x}}_m$, with respect to the ensemble mean $\bar{\underline{\mathbf{x}}}$. In (1), the matrix $\underline{\mathbf{B}}$ is partitioned as in

$$\underline{\mathbf{B}} = \beta_c^2\mathbf{B}_c + \beta_e^2\underline{\mathbf{B}}_e\,, \tag{4}$$

and amounts to a linear combination of a traditional climatological 3D error covariance matrix $\mathbf{B}_c$ with a 4D error covariance $\underline{\mathbf{B}}_e$ formed from the ensemble perturbations $\delta\underline{\mathbf{x}}_m^e$, that is,

$$\underline{\mathbf{X}} = \frac{1}{\sqrt{M-1}}[(\underline{\mathbf{x}}_1 - \bar{\underline{\mathbf{x}}}), (\underline{\mathbf{x}}_2 - \bar{\underline{\mathbf{x}}}), \cdots, (\underline{\mathbf{x}}_M - \bar{\underline{\mathbf{x}}})]\,, \tag{5}$$

so that

$$\underline{\mathbf{B}}_e = (\underline{\mathbf{X}}\underline{\mathbf{X}}^T)\circ\mathbf{C}\,, \tag{6}$$

where $\mathbf{C}$ is a localization covariance matrix typically specified to reduce or wipe out noisy small ensemble correlations from $\underline{\mathbf{X}}\underline{\mathbf{X}}^T$ due to the limited size of the ensemble, while leaving larger correlations unchanged (e.g., Clayton *et al.* 2013). The parameters $\beta_c$ and $\beta_e$ specify the interplay between the climatological

[1] The current implementation of the 4D problem treats $\underline{\alpha}_m$ as a time independent quantity, that is, $\underline{\alpha}_m = \alpha_m\underline{\mathbf{I}}$.

and ensemble background error covariances, respectively. The problem is reset to its traditional 3D-Var configuration, with solution $\delta\mathbf{x}$, when $\beta_c = 1$ and $\beta_e = 0$. Alternatively, a Hybrid 3D-Var configuration is obtained by setting $K = 0$. In this work, the 4D observation error covariance matrix, $\underline{\mathbf{R}}$, is taken to be block diagonal whose blocks are formed from the observation error covariance matrices, $\mathbf{R}_k$ for each sub-interval $k = 0, 1, \cdots, K$ of the assimilation window.

Such extensions from the traditional 3D-Var configuration have been available in GEOS atmospheric data assimilation system (ADAS) since 2013. Hybrid 3D-Var (Version 1.01 of this document; Todling and El Akkraoui 2013), was implemented in the Global Modeling and Assimilation Office (GMAO) Forward Processing (FP) System in mid-2015, followed with an upgrade to Hybrid 4D-EnVar in early 2017. The present document is not meant to provide a comprehensive review of the techniques and works in hybrid data assimilation. For that the reader is referred to the literature: Hamill and Snyder (2000), Lorenc (2003), Wang *et al.* (2007), Lorenc *et al.* (2015), and Fletcher (2017), Chapter 20.

In GEOS ADAS the variational problem of minimizing (1) is solved using the Gridpoint Statistical Interpolation (GSI; Kleist *et al.* 2009b) analysis and the preconditioning strategy of El Akkraoui *et al.* (2013), which is an equivalent alternative to the preconditioning of Derber and Rosati (1989). The climatological background error covariance matrix is implemented as a series of recursive filters producing nearly Gaussian and isotropic correlation functions following Wu *et al.* (2002), and tuned using GEOS forecasts and the NMC-method [Wei Gu contribution; see Rienecker and coauthors (2008)]. In GSI, satellite radiances are processed using the Community Radiative Transfer Model (CRTM; Kleespies *et al.* 2004) and the online variational bias-correction procedure of Derber and Wu (1998)[2]. A normal-mode-based balance constraint term following Kleist *et al.* (2009a) is applied to the climatological part of the increment as well as to the ensemble part of the increment whenever the hybrid analysis is used.

The hybrid-capable version of GEOS ADAS relies on the GEOS global atmospheric general circulation model (AGCM), developed at Goddard. The AGCM includes the hydrostatic finite-volume hydrodynamics of Lin (2004), and recent upgrades to its more advanced cubed-sphere (Putman and Lin 2007). The GEOS AGCM is built under the Earth System Modeling Framework (ESMF) infrastructure of Collins *et al.* (2005), and couples together various physical packages including a modified version of the Relaxed Arakawa-Schubert convective parameterization scheme of Moorthi and Suarez (1992), the catchment-based hydrological model of Koster *et al.* (2000), the multi-layer snow model of Stieglitz *et al.* (2001), and the radiative transfer model of Chou and Suarez (1999). The AGCM also uses the Goddard Ozone Chemistry Aerosol Radiation and Transport (GOCART; Colarco *et al.* 2010).

The 3D version of the GEOS assimilation system uses the incremental analysis update (IAU) procedure of Bloom *et al.* (1996); its 4D counterpart uses a so-called nudged 4D-IAU (N4DIAU) approach, soon to be replaced with a nearest-time 4DIAU approach modulated with a digital filter, as described in Takacs *et al.* (2018). A schematic representation of two consecutive cycles of the IAU-based GEOS ADAS appears in the left panel of Fig. 1; both 3D and 4D flavors of IAU fit the schematic. In the first assimilation cycle, depicted in the panel, 6-hourly observations available around 0000 UTC are combined with background fields derived from an unforced integration of AGCM (represented by the green inverted triangle), to feed into the GSI analysis (purple boxes) which is responsible for producing an incremental correction to the background that forms IAU tendency term(s), which feed into a subsequent 12-hour model integration. In this, the model is forced with the IAU tendency term(s) during the first 6-hours (red triangles), and continues to integrate for at least another 6-hour period when the IAU tendency term(s) is(are) set to zero. The IAU procedure can be thought of as a predictor-corrector scheme, where the first six hours of model integration corresponds to the *corrector* step, and the latter six hours corresponds to the *predictor* step. A complete assimilation cycle encompasses only a 12-hour model integration period, though the model can typically continue integrating beyond twelve hours to complete a short- to mid-range forecast (horizontal orange-dashed lines). Once the first twelve hours of integration is complete, the DAS can proceed to the following assimilation cycle, 0600 UTC in the figure, as long as observations are also available. Throughout this document the cycle just

[2]The satellite bias estimation procedure is in the process of being upgraded to use the enhancements described in Zhu *et al.* (2013).

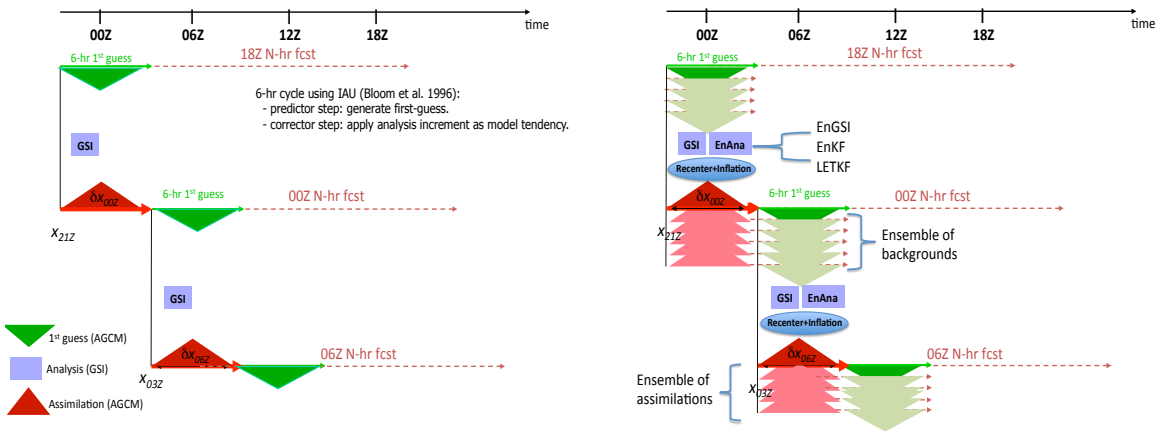


**Figure 1:** Schematic of a traditional assimilation scheme (left) and an ensemble-based assimilation scheme (right) as implemented in GEOS.

described is referred to as the *central* or *deterministic* ADAS.

The process of running GSI and AGCM over and over takes place whether GEOS ADAS is performing traditional 3D-Var or any of its hybrid analysis extensions. The difference between the traditional and hybrid options is that in the latter case an ensemble of background fields is required for GSI to augment its background error covariance formulation, following (4). Furthermore, the differences between having the *central* ADAS perform either Hybrid 3D-Var or Hybrid 4D-EnVar are: (i) GSI requires a higher frequency ensemble of background fields in the latter than in the former; and (ii) the corrector step of IAU changes from its traditional 3D flavor (Bloom *et al.* (1996)) to one of its 4D flavours (Takacs *et al.* 2018). A third difference between the 3D and 4D options in GEOS ADAS relates to how GSI employs a first-guess at the appropriate time (FGAT; Massart *et al.* 2010) strategy to calculate observation residuals, such as $\underline{\mathbf{y}} - \underline{\mathbf{y}}^o$ appearing in (1): in 3D, GSI-FGAT uses 3-hourly backgrounds; in 4D, GSI-FGAT uses 1-hourly backgrounds.

As mentioned above, a *hybrid* version of the deterministic cycle depicted in the left panel of Fig. 1 requires the availability of an ensemble of background fields to make up the ensemble background error covariance $\underline{\mathbf{B}}_e$. This, in turn, entails an ensemble of model integrations that must start from an ensemble of "initial conditions" (analyses; or IAU tendencies). This is the responsibility of GEOS Ensemble ADAS (EnADAS). At least three options exist within GEOS EnADAS to generate an ensemble of analyses. The standard option follows Whitaker *et al.* (2008) and relies on the ensemble Kalman filter (EnKF) software of J. S. Whitaker, from NOAA/ESRL; specifically, its so-called ensemble square-root filter (EnSRF) option (also used in the NCEP operational global data assimilation system). Alternatively, GEOS EnADAS can generate an ensemble of GSI analyses to enable an Ensemble of Data Assimilations (EDA) strategy in the spirit of ECMWF's EDA (Isaksen *et al.* 2010). The necessary details for the construction of a reliable EDA within GEOS EnADAS are, however, not fully complete and this option is not discussed further. Lastly, a simplified ensemble generation procedure, referred to as the Filter-free Ensemble (FFEn) scheme, is also available in GEOS EnADAS. In this procedure, ensemble analyses are created by simply inflating the central (hybrid) analysis with NMC-like perturbations[3]. Regardless of the ensemble of analyses scheme used, once an ensemble of analyses is available, a corresponding set of background fields is generated through IAU-based AGCM integrations of the members of the ensemble.

The augmentation of the GEOS assimilation cycle when an EnADAS is required is illustrated in the right

[3]These perturbations are part of what is used to employ the so-called NMC method introduced by Parrish and Derber (1992) to estimate climatological background error covariances.

panel of Fig. 1. The schematic for the EnADAS is essentially a reproduction of that of the deterministic ADAS, with minor differences. Similarly to the central ADAS, once observations and (now) an ensemble of background fields are available, any one of the ensemble analysis options (EnAna; right-placed, purple boxes) generates an ensemble of analyses, which are turned into an ensemble of IAU tendencies used to feed into a corrector (red triangles) and predictor (green inverted triangles) sequence of 12-hour ensemble model integrations. Note that presently, regardless of whether the central ADAS performs a Hybrid 3D-Var or a Hybrid 4D-EnVar, the EnADAS performs only a 3D analysis of the members (either with the EnSRF or the FFEn scheme); thus the ensemble of model integrations employs a 3D IAU procedure (moduled by a digital filter).

Another schematic view of the hybrid ADAS implementation appears in Fig. 2; this schematic highlights additional subtleties of the interplay beween the central and ensemble systems. Here, the two data assimilation systems are shown to run parallel to each other, as represented by the two grey-shaded blocks. The top grey-shaded block corresponds to the deterministic hybrid assimilation component of the system: given observations (OBS), corresponding bias correction (OBC), and background fields (BKG), each cycle entails running a hybrid GSI that generates analysis increments to be then taken in by the GEOS AGCM through an IAU-like procedure. The bottom grey-shaded block corresponds to the ensemble ADAS which takes observations, corresponding bias correction, and an *ensemble* of background fields to calculate an *ensemble* of observation-minus-background (OMB) residuals which are fed into an ensemble analysis procedure (e.g., EnSRF). This in turn produces an *ensemble* of IAU-forcing terms that are used to integrate an *ensemble* of GEOS AGCMs responsible for generating the set of ensemble backgrounds required for the subsequent cycle of the ensemble assimilation system and the corresponding hybrid analysis of the deterministic system.

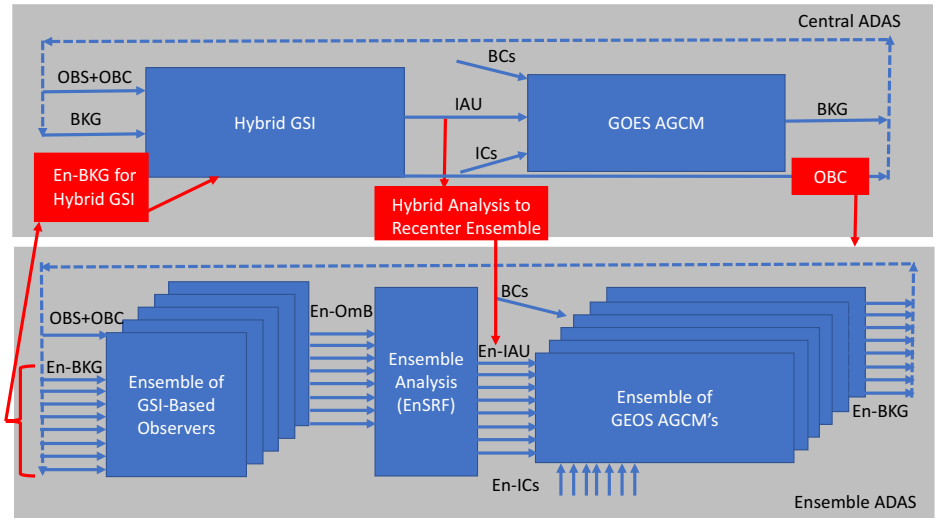


**Figure 2:** Schematic of the GMAO Hybrid Ensemble-Variational Data Assimilation System. The two grey-shaded blocks represent the central (variational) ADAS (top) and the ensemble ADAS (bottom). Both are IAU-based systems whose analyses require background fields (Bkg), observations (Obs) and estimates of observation bias correction terms (OBC), and AGCM integrations that require initial conditions (ICs) and boundary conditions (BCs). See text for complete description of figure.

This second view of GEOS Hybrid EnADAS emphasizes the two-way feedback between its *deterministic* and *ensemble* components. Specifically, the red arrows and boxes illustrate how the two systems communicate with each other. Naturally, a one-way feedback is established from the *ensemble* to the *hybrid* analysis with the latter requiring an ensemble of background fields (left-most BKG red box and arrow). A two-way feedback mechanism is established by the *ensemble* requiring both satellite bias correction coefficients (right-most OBC red box and arrow) and hybrid analysis (middle red box) from the central ADAS.

The first ensures the observations used by the ensemble observers [$M$ instances of $\underline{\mathbf{y}}^o - \underline{H}(\underline{\mathbf{x}}^b)$] are the same as those seen by the hybrid (central) analysis; the second ensures that the ensemble predictions stay close to the central ADAS predictions. The latter is accomplished by replacing the ensemble mean of the analysed members with the hybrid analysis. This two-way feedback can be decoupled by having the ensemble analysis estimate the biases in the observations it uses, and by skipping the re-centering step. Given that it would be hard to have a consistent observing system between the two assimilation systems if bias correction was treated separately, the work here never investigates the effect of removing this component of the feedback. However, some work has been done to assess the effect of re-centering (see below).

The re-centering step is usually accompanied, or followed, by inflation of the ensemble analysis members (see blue boxes in Fig. 1). As mentioned above, re-centering is done in order to align the ensemble of analyses with the hybrid GSI analysis and avoid a possible divergence of the ensemble. Additive inflation is applied to compensate for lack of better ways to represent model error. It should be noted that in case of the EnSRF, an attempt is also made to account for sampling errors by applying a multiplicative relaxation to prior inflating factor to the analyses[4]. For more on the effects of inflation see the works of Hamill and Whitaker (2005), Whitaker *et al.* (2008), Charron *et al.* (2010), and Bowler *et al.* (2017).

| | Experiments | | Near-Real-Time System | |
|---|---|---|---|---|
| | Central ADAS | 32-mem EnADAS | Central | 32-mem EnADAS |
| AGCM | C360 (25 km) | C90 (100 km) | C720 (12.5 km) | C180 (50 km) |
| GSI Observer | 576×361 | 288×181 | 1152×721 | 576×361 |
| GSI Solver | 576×361 | N/A | 1152×721 | N/A |
| EnSRF Solver | N/A | 288×181 | N/A | 576×361 |

Table 1: Two relevant configurations of the Hybrid 4D-EnVar GMAO ADAS: columns under *Experiments* refer to configuration corresponding to what is used for typical experiments performed for most testing and science study purposes; columns under *Near-Real-Time* refer to configuration corresponding to present (since January 2017) settings of the GMAO Forward-Processing System. Resolution is shown in degrees when referring to components of the system operating on a regular latitude-longitude grid; resolution of components operating on the cubed-grid are indicated with a number preceeded with the character C; all components work with 72 levels in the vertical.

What follows in this Section is an update of the results presented in the original documentation that accompanied the first release of the GEOS Hybrid 3D-Var Ensemble-Variational Atmospheric Data Assimilation System (EVADAS) to the GMAO Quasi-Operational group in 2013. The sub-sections below give a brief summary of tests and experiments done to upgrade GMAO FP System from its Hybrid 3D-Var to a Hybrid 4D-EnVar configuration. The upgrade of FP occurred in January 2017. The configuration of GEOS Hybrid EVADAS presently running as the GMAO near-real time system is shown in Table 1 (see details in what follows). As stated earlier, this document is not meant to be in any way a comprehensive review of data assimilation techniques and in many ways it assumes considerable familiarity with those concepts. The reader is referred to the works of Lewis *et al.* (2006), Bannister (2017), and Fletcher (2017) for academic exposition of data assimilation concepts. Readers familiar with the state of science in GEOS ADAS are encouraged to skip to the next section where details of use and options for actually conducting experiments with GEOS Hybrid EVADAS are presented.

[4]This is an option within the EnKF software. In GEOS ADAS, a single program is responsible for both re-centering and additively inflating the member analyses (see Secs. 7.1 and 7.2).

## 1.2 Brief summary of main components in GEOS Hybrid EVADAS

The illustrations in this section derive from a variety of experiments conducted with GEOS Hybrid EVADAS. Some come from single, non-cycling, analysis with the EnSRF and hybrid GSI, while others come from fully-cycled assimilation experiments. The intention is to give not only an overall update of the state of GEOS Hybrid ADAS, but also to point out what possible changes, enhancements, and upgrades are likely to be expected in upcoming releases. Unless indicated otherwise, the evaluations discussed here correspond to a system configuration shown in the column under Experiments in Table 1; a comprehensive report on the performance of the 12.5 km FP System is to appear elsewhere.

### 1.2.1 Configurations of GSI Observer and EnSRF

The GMAO implementation of the ensemble data assimilation follows Whitaker *et al.* (2008), and in this sense, it is similar but not identical to that of NCEP's. GMAO and NCEP global assimilation systems share the same GSI and EnKF base software, but the actual configuration of their corresponding systems is substantially different. Not only do the grid specifications differ between GMAO and NCEP background fields (former operates on a regular grid; latter on a Gaussian grid; see also Table 1), but also choices made with respect to the treatment of observations (e.g., satellite channel selection; data thinning; quality control), GSI minimization options, initialization strategies, and numerous other details make both analysis systems rather different in actuality.

Differences are also present when it comes to the configuration of the EnSRF and related (GSI) observer. At the time of this writing, NCEP thins satellite observations rather aggressively in the observer step related to the ensemble analysis, as opposed to the GSI corresponding step; GMAO, on the other hand, thins observations equally in both central and ensemble observers. Combined with the settings in the ensemble analysis procedure, the choice of observer thinning strategies has consequences on what observations end up actually being used by the EnSRF. Furthermore, there are at least three ways of handling the observations under the EnSRF sequential algorithm, where observations are assimilated one at a time: (i) assimilate observations in the order they are read; (ii) randomly sort observations before assimilating; and (iii) assimilate observations in order of increasing (local) analysis to background error ratio. This last amounts to a scheme that tries to assimilate first the observations impacting the analysis the most; this is sometimes referred to as a degrees of freedom for signal (DFS) selection criteria since it resembles a DFS-like diagnostic (see Lupu *et al.* 2011). Scheme (iii) can be rather costly as it requires multiple passes over the observations. In addition to these three observation handling options, a parameter controlling the degree of "thinning" of the observations is implemented such that only observations that decrease error variance by a specified percentual amount are assimilated. This setting is also linked to the specified background error localization scales. Without too much detail, one can get the EnSRF to assimilate more observations by either tightening the localization scales or increasing the variance percentual reduction toward one hundred (when all observations are assimilated[5]).

An illustration of what happens to the observation count in the EnSRF when the various options of observation handling above are made is given in Figure 3. The example here is extracted for an experiment done at 1200 UTC, 30 November 2016, when the ensemble mean observer using the GMAO thinning selects 3,887,557 observations. In this case, the thinning in the EnSRF observers is consistent with that of the hybrid GSI (GMAO) configuration. With this, three cases are considered as selection criteria for offline runs with the EnSRF: (a) assimilation of observations in the order they are read, and only those reducing the error variance by at least 2.0%; (b) assimilation of observations based on DFS, and only those reducing error variance by at least 2.0%; and (c) assimilation of as many observations as possible, regardless of their level of contribution to error variance reduction. The figure shows observation count, for each observation type, scaled by the number of observations accepted by the ensemble mean observer for that corresponding type. As we have mentioned earlier, even in case (c), when the EnSRF is set to use as many observations as

[5]The software is actually protected to prevent roundoff errors and assimilation of redundant observations so that it never allows all observations to be assimilated, but very close to that.

possible, it does not quite use all the observations; this difference is seen in the figure where the "Total" green bar does not quite reach 100%. Examining the figure, we see that when the EnSRF uses observations in the order they are read, and only admits those contributing to the specified error reduction level, it ends up wih a very small percentage of radiance observations; with the DFS criteria the percentage of radiances used increases somewhat, but it is still dramatically lower than the percentage of conventional observations, ozone, and GPS. Indeed, GPS is the observing system most consistently treated among all three of these observation handling options. Though the GMAO FP system currently uses option (a), we are in the process of considering its replacement with option (c)[6]. Allowing the EnSRF to use as many observations as possible makes the EnSRF analysis more consistent with the central hybrid GSI analysis. In the case supporting the experiments just discussed, the central GSI observer takes in a total of 3,889,861 observations, which is a number very close to what the ensemble mean observer takes in.

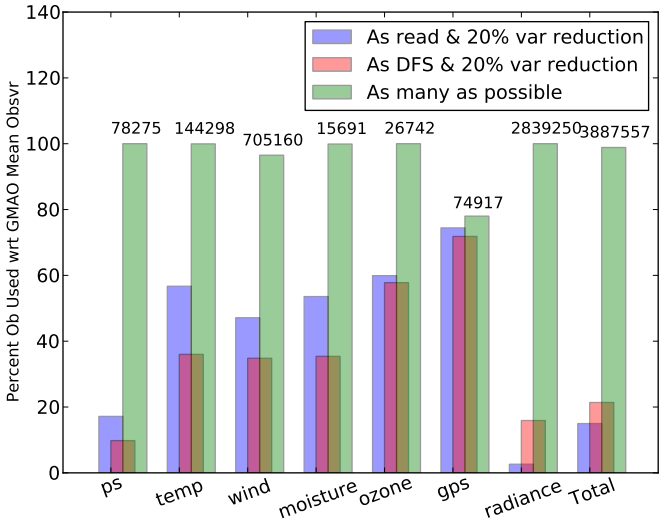


**Figure 3:** Count of observations used in EnSRF given different choices in observer and/or EnKF observations handling options.

There is a myriad of consequences and implications related to the choice of these EnSRF observation handling options. The most obvious one is of a practical consequence: the computational performance of the EnSRF is highly dependent on these options. In the settings for the experiments discussed in Fig. 3 the most computationally burdensome is the DFS-based option (b). This is not surprising given the multiple passes over the observations required by this method. Option (a) is the most computationally efficient, and indeed it was chosen by NCEP exactly for this reason; GMAO followed the same choice. However, when the EnSRF uses a lot fewer observations than the hybrid GSI does, as in option (a), the two analyses are bound to be considerably different. A consequence, for example, is that analysis re-centering must play considerably different roles when faced with each one of these options. The effect of re-centering is illustrated in the following subsection.

Before closing the discussion related to options of handling the observing system in the EnSRF, we show how evidence of consistency, or lack thereof, between the hybrid (central) and ensemble analyses can be seen in other diagnostics. One useful diagnostic to ascertain the level of tuning in a data assimilation system is the diagnostic provided by the trace value of the product of the linearized observation operator, $\mathbf{H}$, and the Kalman gain matrix, $\mathbf{K}$. These matrices are not explicitly available in practice, but it can be shown

[6]We should point out that it was not until recently (circa mid-2017), while working with Fabio Rodrigues Diniz, a CPTEC/Brazil graduate student visiting GMAO, that we came to realize that the EnSRF software did not properly flag observations actually assimilated, giving the impression that a lot more observations participated in the minimization than actually do.

(e.g., Eyre 2016) that an approximation for Tr(**HK**), scaled by the number of observations, is given by

$$\mathrm{Tr}(\mathbf{HK})/p \approx 1 - [E(J_{of})/E(J_{oi})]^{1/2} \tag{7}$$

where $E(\bullet)$ represents an ensemble average operator, and $J_{oi}$ and $J_{of}$ represent the evaluation of the second term on the rhs of the cost function in (1) at the initial and final iterations of the underlying minimization procedure, respectively. Whether in the variational framework or its dual counterpart (e.g., Courtier 1997) the terms $J_{of}$ and $J_{oi}$ are either immediately available or easily obtainable. Eyre (2016), for example, reports that in the U. K. Met Office global 4D-Var system (ca. 2016) the value of $\mathrm{Tr}(\mathbf{HK})/p$ was about 0.2.

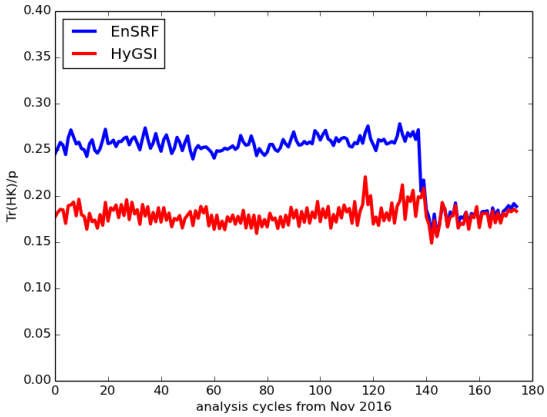


**Figure 4:** Time series of Tr(HK)/p for both hybrid GSI (red curve) and EnSRF (blue). Results are displayed for every cycle from 0000 UTC 01 November 2016 to mid December 2016, for a typical experiment with GEOS Hybrid EVADAS. Up to cycle 134 the EnSRF is configured to assimilate observations in the order they are read, and only takes contribution from observations reducing errors at least 2.5%; from cycle 135 onward, the EnSRF is re-configured to assimilate as many observations as possible regardless of how much error reduction they amount to.

Figure 4 shows a time series of this value calculated from a typical hybrid experiment performed with the GMAO Hybrid 4D-EnVar system. Results here cover almost 50 days of assimilation and show the value of $\mathrm{Tr}(\mathbf{HK})/p$ for the central hybrid GSI (red curve) to be around 0.18, not too far from the value reported in Eyre (2016). The corresponding time series calculated from the observation residuals of the EnSRF is shown as the blue curve. Up to cycle 134, the EnSRF is configured to assimilate observations in the order they are read and to take contributions from only those observations reducing the error by at least 2.5% (roughly option (a) above). From cycle 135 onward the EnSRF is reconfigured to assimilate as many observations as possible, of those selected by the mean observer (consistent with option (c) above). It is only when the EnSRF is configured to use almost as many observations as the hybrid GSI that its diagnostic resembles the corresponding diagnostic from the hybrid GSI. Throughout the time series in this figure only residuals *actually used* in either the hybrid GSI or the EnSRF analyses participate in the calculation of the trace diagnostic[7]. It is important to note that reconfiguration of the EnSRF to use all observations leads to a slight reduction in global ensemble spread (not shown).

[7] As it turns out, when the calculations in the (original) EnSRF software are performed using *all* observations selected by the ensemble mean observer, regardless of whether they are actually used in the EnSRF minimization or not, the diagnostic incidentally shows numbers not too distant from those of the hybrid GSI - averaging around 0.15 (not shown here). This can easily mislead us into thinking that the EnSRF analyses are in sync with the hybrid analyses whereas, in truth, when the calculation is done correctly this is clearly seen not to be the case as the illustration in Fig. 4 shows.

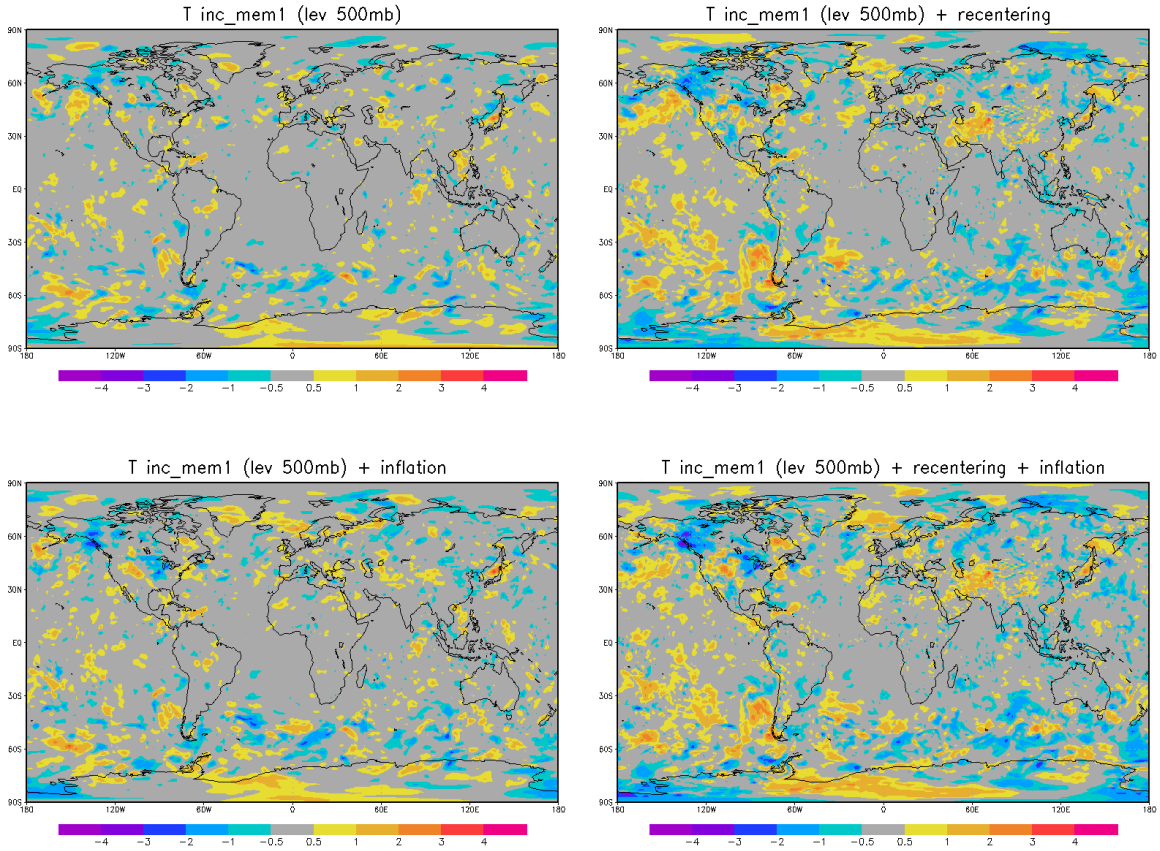


**Figure 5:** Illustration of contribution from each step taking place after the EnSRF ensemble of analyses are generated. The panels show 500 hPa virtual temperature: analysis increment for a given ensemble member (top left); effect of re-centering this particular member about the central GSI analysis (top right); effect of applying additive inflation to the same member analysis with a coefficient of 0.25 (bottom left); and the resulting increment after both re-centering and additive inflation are applied (bottom right).

### 1.2.2 Re-centering and inflation

Re-centering and inflation have been key components of the GMAO hybrid implementations. In GEOS ADAS re-centering and inflation amount to an overwrite of the $m$-th member EnSRF analysis $\mathbf{x}_m^a$ as in

$$\mathbf{x}_m^{ra} = \mathbf{x}_m^a - \bar{\mathbf{x}}^a + \mathbf{T}\mathbf{x}^a + \alpha\mathbf{T}'\delta\mathbf{p}_m \tag{8}$$

where $\bar{\mathbf{x}}^a$ is the original ensemble mean analysis at a given time, $\mathbf{x}^a$ is the central analysis at the same time, and $\mathbf{p}_m$ is a perturbation vector for member $m$. Since, as seen in Table 1, the central ADAS analysis is typically produced at a resolution higher than that of the ensemble members, the central analysis needs to be converted down to the resolution of the members. The related operation is represented by the matrix $\mathbf{T}$ in the expression above. As it turns out, this conversion is better referred to as a re-mapping procedure since not only the resolution is changed, but the topographic differences in the high resolution central analysis are made consistent with the topography seen by any given member of the ensemble. The additive perturbations $\mathbf{p}_m$ are selected randomly from a database of 48-minus-24 hour forecast differences created from a little over one year of forecasts from the GMAO near-real-time system; for a given date and time, only perturbations falling within a range of 45-days around the current time are selected from the database. In other words, the perturbations used in (8) are seasonally consistent. This database of perturbations is the same as that used to estimate the climatological background error covariance matrix, $\mathbf{B}_c$ through the NMC-method (e.g.,

Parrish and Derber 1992 and Bannister 2008). Depending on the resolution of the ensemble members, these perturbations[8] need to go through an interpolation operator, represented here as $\mathbf{T}'$. The parameter $\alpha$ controls the magnitude of the additive perturbations and is typically set to 0.35 in the GMAO system. It is useful to notice that it is rather simple to treat (8) in its alternative incremental form by simply subtracting the ensemble background of the member in question from both sides of the equation.

Figure 5 shows the effect of re-centering and additive inflation when applied to a given member of an EnSRF analysis. Each panel in the figure shows the individual contribution to the increment: EnSRF only (top left); EnSRF plus re-centering (top right); and EnSRF plus additive inflation (bottom left); the final increment when both re-centering and inflation have been applied to the EnSRF increment appears in the bottom right panel. If care is not taken, either one of these two operations might do more than the EnSRF itself. When the underlying ensemble assimilation system is not well tuned, it is possible that increments due to re-centering become so large as to wipe out the EnSRF increments; similarly, when additive inflation is too large, it can easily overwhelm increments from the EnSRF. An adequate balance between tuning the EnSRF to have a mean state that is reasonably close to the central analysis and the magnitude of additive inflation must be reached to obtain an effective total increment for each member of the ensemble. In general, however, when all is tuned reasonably well, re-centering still provides a contribution nearly as large as that obtained from the EnSRF analyses.

Whether due to the relatively small number of ensemble members or simply to the nature of the hybrid procedure, lack of re-centering of the members around the central analysis has undesirable consequences in GEOS Hybrid ADAS. An illustration of the effect of re-centering is given in Fig. 6 where the time evolution of the *background* sea level pressure field is shown at four consecutive synoptic times from 0000 UTC 24 December 2016 during the evolution of Typhoon Nock-ten in the West Pacific. The left column is from an experiment in which the ensemble analyses are not re-centered around the central ADAS analyses, and the right column is from an experiment using the re-centering strategy. Contours of sea level pressure are shown for the "high-resolution" (C360) central background (red curves) as well as for the low-resolution (C90) ensemble mean background (black curves); shaded areas correspond to ensemble spread. The GMAO Hybrid 4D-EnVar system is used for these experiments, and in these, no cyclone relocation strategy (see Kleist 2011) is used, as is normally the case in our Hybrid 3D-Var system configuration. The GMAO ensemble members are never relocated, regardless of what hybrid variational flavor is employed by the central ADAS. It is clear from the figures that when the ensemble is not re-centered the member backgrounds have considerable freedom to locate the storm at positions not quite consistent (whether correct or not) with the position predicted by the central ADAS. It is noticeable that in the non-re-centered case (left), on average, the ensemble positions the storm at different location than where the central ADAS does. As the storm intensifies this becomes evident even in the ensemble spread, with results from a re-centered system (right) having more compact spread than when no re-centering is applied. The lack of re-centering is even noticeable farther away from the storm where the ensemble mean stream lines of sea level pressure are not quite aligned with those from the central ADAS as opposed to their fairly nice alignment in the re-centered case (right).

### 1.2.3 Non-cycling hybrid analysis

When an ensemble of backgrounds is used in a hybrid (central) GSI analysis, it is useful to examine how the analysis increment changes with respect to its non-hybrid counterpart. Figure 7 provides an illustration of the change in the analysis increment, measured in total energy units, for an analysis calculated at a single synoptic time using: (i) a regular 3D-Var GSI, with only the climatological background error covariance matrix (left); (ii) a 3D-Var GSI with a background error covariance matrix that is fully determined from a 32-member ensemble (center); and (iii) a Hybrid 3D-Var GSI when 50% of background error covariance matrix comes from the ensemble and the remaining 50% comes from its regular climatological background error

[8]The NMC-perturbations in the GMAO database are generated on regular latitude-longitude 721x1152 grid, corresponding roughly to 25 km.

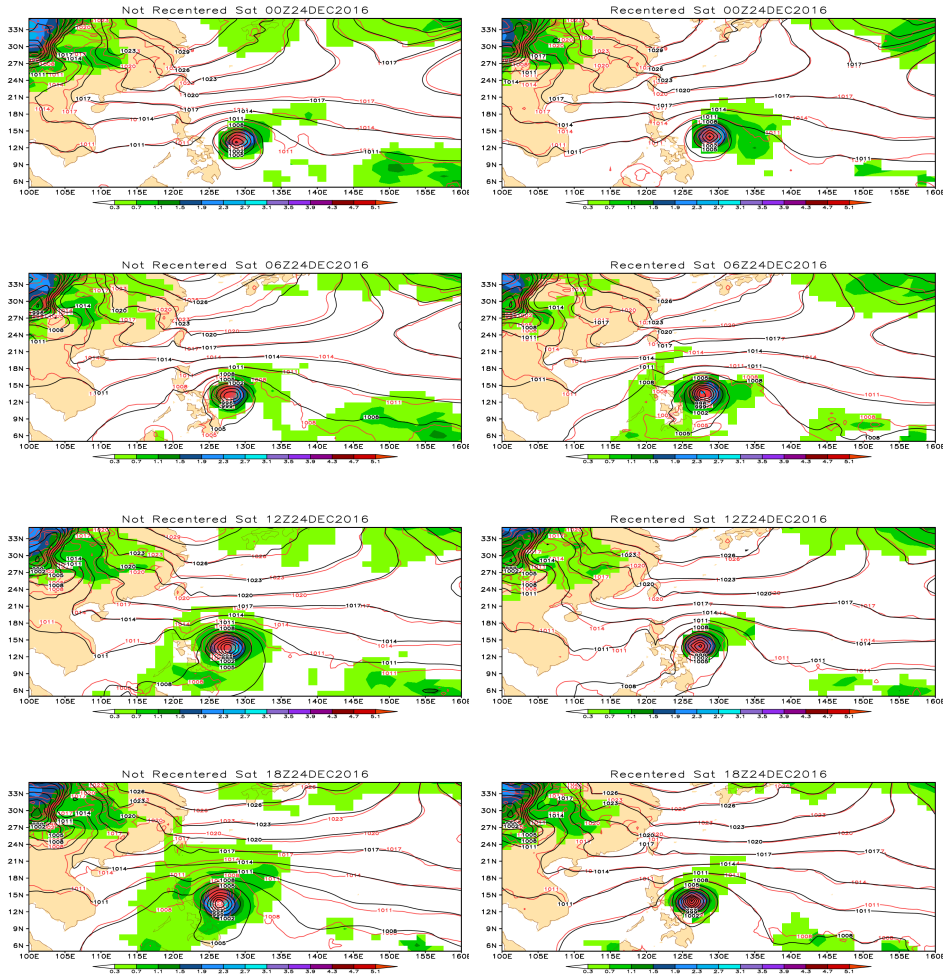


**Figure 6:** Six-hour evolution of Typhoon Nock-ten in the West Pacific just before reaching Category 5 on 25 December 2016. Sea level pressure is shown by the red curves for background fields at times indicated in each panel, with the corresponding ensemble mean sea level pressure shown by the black curves. The shaded areas show the ensemble spread in each case. The panels on the left are from a cycling experiment in which the ensemble analyses are not re-centered around the central ADAS analyses; the panels on the right are from an experiment in which the member analyses are re-centered.

covariance matrix (right). The ensemble-only case (center) shows considerably more activity in the tropics than the climatological-only case (left); the resulting hybrid (right) increment shows slight but noticeable energy increase in the mid-tropospheric and low-stratospheric levels — a little less energy seems to be present along the Southern tropospheric jet in the ensemble (center) when compared with the climatological case (left), with the resulting hybrid retaining the energy in this region (right).

Another aspect of relevance when introducing hybrid analyses as replacements for regular 3D-Var analyses relates to how balance gets affected. In its 3D-Var configuration, GSI has the capability of applying a tangent linear normal mode constraint (TLNMC) to its increments (see Kleist *et al.* 2009a). In its more general hybrid form (3) the constraint can be applied to either term of the total increment (see Kleist 2012). Figure 8 shows two illustrations of the result of balancing the increment in various configurations of GSI. The panel on the left shows the total cost function during the iterations of the GSI minimization when using: traditional 3D-Var without TLNMC (black curve); traditional 3D-Var with TLNMC (red curve); Hybrid 3D-Var with TLNMC applied only to the climatological part of the increment (green); and Hybrid 3D-Var when TLNMC is applied to the full increment. The behavior is typical of what happens when adding constraints to the analysis, that is, with balance, the cost settles a little higher than when no constraint is applied (see

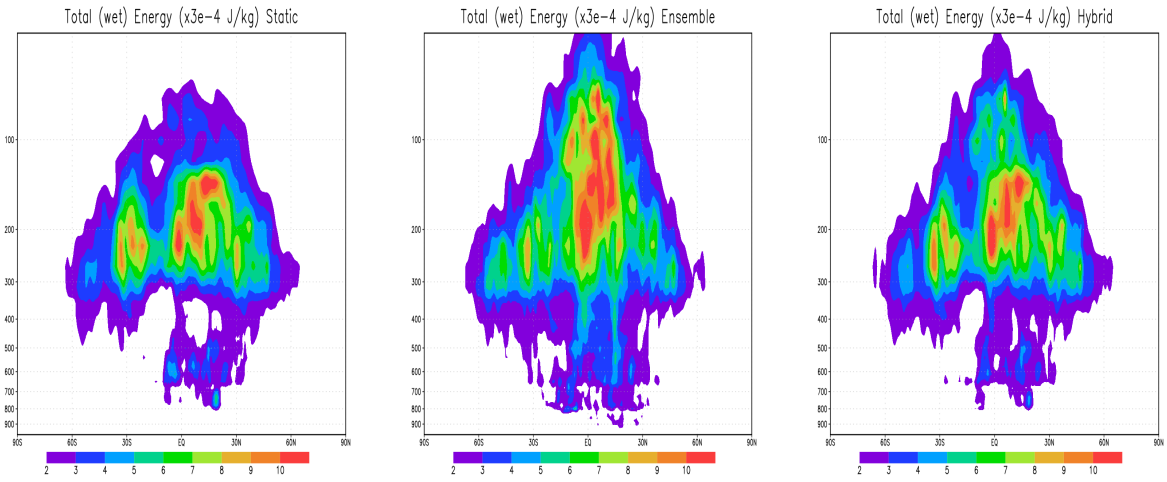


**Figure 7:** Zonal mean analysis increment, in total wet energy (J/kg) norm, using a standard 3D-Var (left), a 3D-Var when the background error covariances are fully determined by the ensemble (center), and a Hybrid 3D-Var when the covariances are a 50% weighted sum of the climatological- and ensemble-derived background error covariances (right).

also Gauthier and Thépaut 2001). Introduction of the hybrid feature reduces the cost, in the case shown in the figure, as compared to the climatological-balanced configuration; this is particularly noticeable in the first outer minimization (first 100 iterations; compare green and blue curves with red curve, respectively). The combination of hybrid and balance constraint is such that the cost ends up at levels comparable to when traditional 3D-Var runs with balance constraint.

The real measure of improved balance is displayed in the right panel of Fig. 8 where the spectra of the vertically integrated mass-wind divergence increment is shown for the same four configurations above. The curves show clearly that TLNMC brings in considerable improvement in balance when applied to traditional 3D-Var (the color scheme is preserved; compare black and red curves). It is also clear from the figure that applying TLNMC only to the climatological part of the increment when Hybrid 3D-Var is used is rather troublesome (green curve). This is natural since there is nothing to guarantee the ensemble contribution to the increment, through its background error covariance matrix $\mathbf{B}_e$, to be balanced in any way; TLNMC must be applied to the full increment (blue curve) for balance to be acceptable in a hybrid configuration. However, this latter case is not completely perfect since some power in the spectrum still remains for large wave numbers which would best be reduced. As pointed out by Kleist (2012, see Figure 4.2 on page 108 in that work), this is a consequence of the dual-resolution aspect of the hybrid analysis and some aliasing of the winds; in the example shown here, the ensemble is generated at half the resolution of the climatological background error covariance (that is, a 1-degree ensemble for a 0.5-degree analysis). It is possible to use scale-dependent weights to reduce some of the aliasing issues Kleist (2012, see Fig. 4.4, in that work). In the case of GEOS Hybrid ADAS, the default is to apply TLNMC to the full increment. Similar considerations apply to the 4D-EnVar and Hybrid 4D-EnVar configurations of GSI. The overall message is similar and the default in the 4D options is to apply TLNMC to all time slots of the 4D increment.

In 4D settings, an additional concern arises since in such cases, there is a possibility for the solution of the minimization problem to overfit the observations (e.g., see Trémolet 2007b). In upgrading GEOS ADAS from Hybrid 3D-Var to Hybrid 4D-EnVar a number of tests were performed in non-cycled and short-cycled cases to determine an adequate number of iterations for the hybrid GSI minimization. During this test phase, we have also looked into the possibility of using, what we refer to as, a *legitimate* outer-loop strategy. In a traditional implementation of incremental 4D-Var such as that of Rabier *et al.* (2000), when tangent linear and adjoint models of the underlying nonlinear model are available, an outer-loop strategy is implementated such that successive linearizations of a cost function similar to that in (1) are minimized

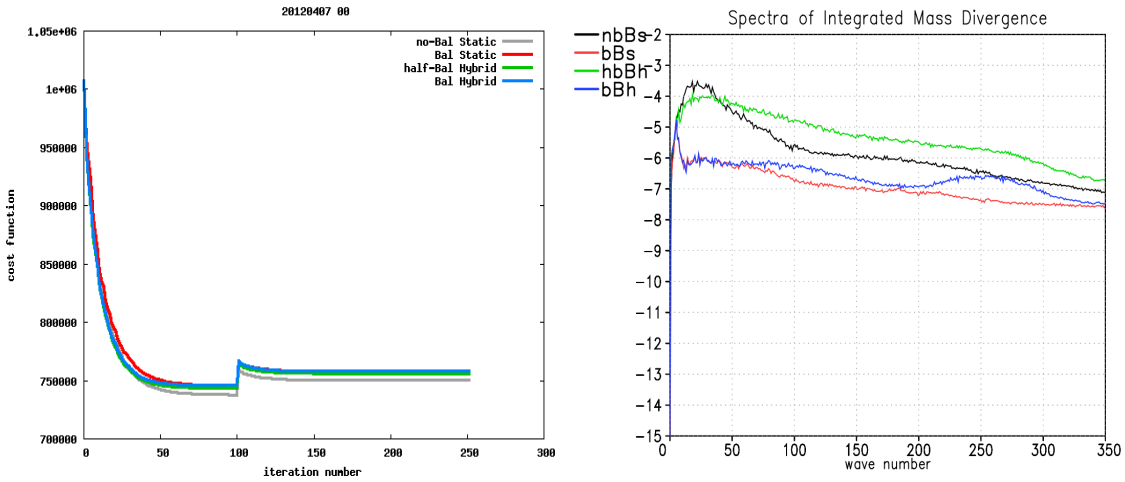


**Figure 8:** The panel on the left shows the total cost function as it changes during the iterations of the GSI minimization; all cases are calculated for the same synoptic time but GSI is configured as follows: climatological (non-hybrid) 3D-Var without balance constraint (black curve); (non-hybrid) 3D-Var with TLNMC balance constraint (red curve); Hybrid 3D-Var without balance constraint applied to hybrid part of increment (green curve); and Hybrid 3D-Var with balance constraint applied to full increment (blue curve). The panel on the right shows the integrated mass-wind divergence spectra of the analysis increment as a function of wave number for the same four configurations; color scheme of curves is as in panel on the left.

after each nonlinear model integration in an incremental strategy to solve the global minimization problem associated with the corresponding version of (1). Most typical of implementations elsewhere, including those using 3D-Var and Hybrid 3D-Var, is a somewhat mixed strategy, referred to as middle-loop, when the nonlinear model is integrated only once (per assimilation time window), but the nonlinear observation operator is successively linearized at will. A schematic configuration of these two strategies is shown in Fig. 9. The top panel illustrates the most common implemetation of a 4D-IAU-based assimilation procedure using a single outer-loop and incorporating a middle-loop strategy; and the bottom panel illustrates the implementation of 4D-IAU-based strategy using two legitimate outer-loops. These implementations are, in principle, valid just as well when the linearized model and its adjoint are replaced with the 4D-perturbation model generated from a hybrid ensemble configuration.

It is illustrative to examine how the cost function minimizes as these strategies are implemented. Figure 10 summarizes results from a study performed with GEOS Hybrid EVADAS. The panel on the right corresponds to experiments performed with the low resolution configuration laid out in Table 1, and results from experiments using the high resolution configuration in the table are shown on the left panel. In low resolution experiments it is possible to have multiple outer loops, each with many iterations in the inner loops[9]. All curves on the left panel correspond to some implementation of Hybrid 4D-EnVar: six legitimate outer loops (red curve), six middle-loops (blue curve), and three middle-loops (green curve). The curves show the value of the cost function (1) as the minimizations progress. The number of iterations in the inner minimizations is chosen to facilitate comparison between the middle-loop and legitimate outer-loop strategies. The first noticeable result is how a middle-loop strategy provides misleading information for how the guess actually fits the observations. For example, at iteration 51, the middle loop strategy re-linearizes the observation operator using an updated guess that simply corresponds to a linear update of the same guess available at iteration 1; on the other hand, the legitimate outer-loop strategy re-linearlizes the same operator based on guess fields coming from a re-integration of the model (see bottom panel of Fig. 9). The cost function calculated for this latter case shows how the model fields *truly* fit the observations; the cost calculation

[9]Notice that the "low" resolution here is not that coarse, and having six outer loops as in one of the illustrations in the figure is only computationally viable for test purposes.

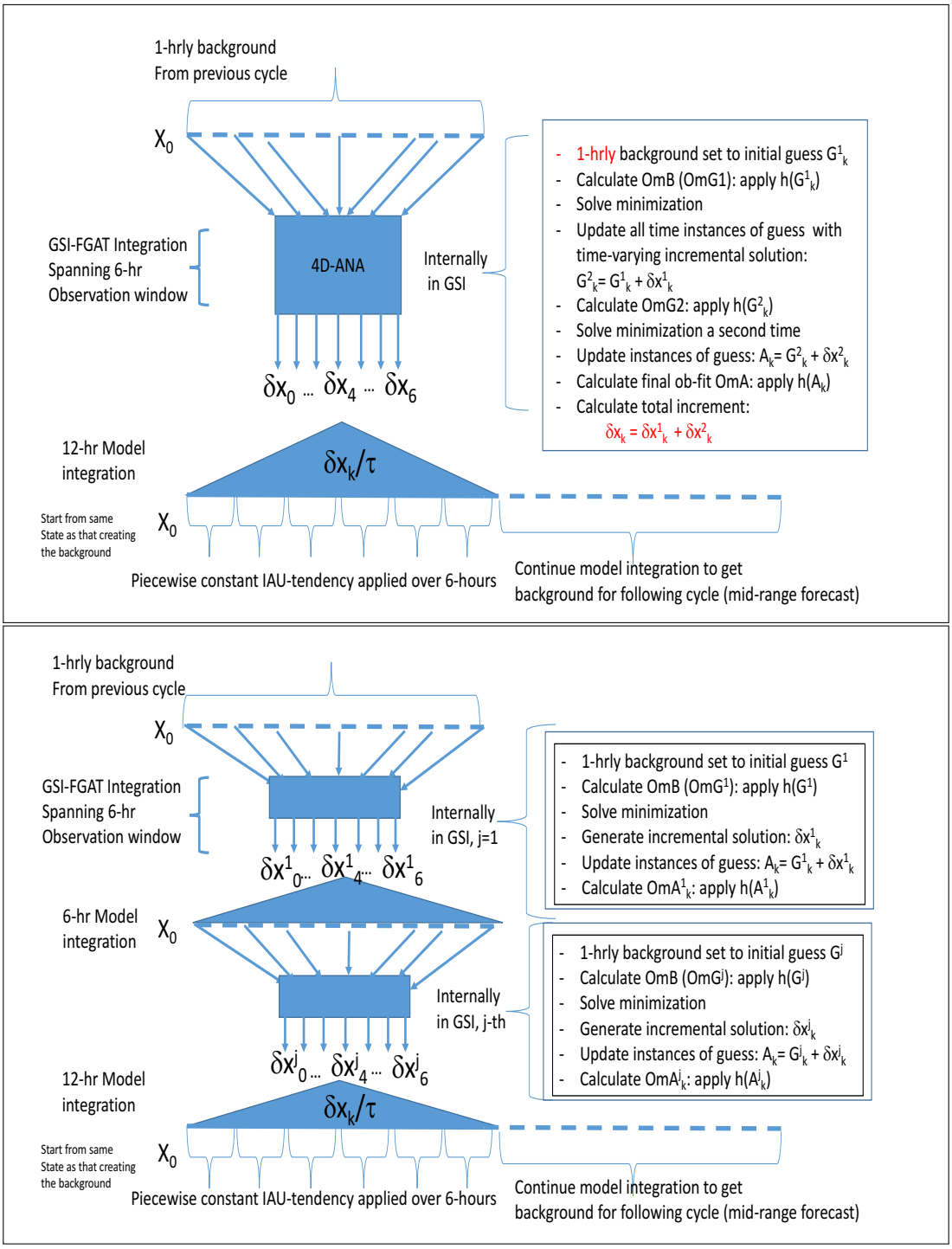


**Figure 9:** Illustration of implementations of middle-loop and legitimate outer-loop strategies for a first-guess at appropriate time (FGAT), 4D-IAU approach. The top panel illustrates a middle-loop strategy using hourly first guesses; the panel on the bottom illustrates a two outer-loop strategy also using hourly first-guesses. The top solid blue line in both panels represent the free (unforced) running model responsible for providing the Hybrid 4D-EnVar GSI (squared/rectangular boxes) with hourly guess fields. The triangles represent the model integrations with the 4D-IAU forcing derived from the incremental GSI minimization. In the case of a legitimate outer loop, both the GSI minimization and the 4D-IAU forced model integrations are performed multiple times (twice in the illustration here).

from the middle-loop strategy cannot know how the model reacts to the corrections due to the first 50 iterations. The same continues to happen as the iterations proceed. Even at iteration 250, when either one of the middle-loop configurations seems fully converged, the outer loop strategy indicates there being still room for error reduction, and that the model does not fit the observations as well as the middle-loop strategy seems to suggest. It must be pointed out that the inner loops above use a simple, so-called, **B**-preconditioning without benefiting from accelerated Hessian preconditioning procedures using a Lanczos algorithm (e.g., Derber and Rosati 1989, Fisher 1998, El Akkraoui *et al.* 2013, and Gürol *et al.* 2014). A Hessian-based Lanczos-preconditioning is available in GSI and future experimentation and actual implementations are expected to exercise it (e.g., Trémolet 2008).

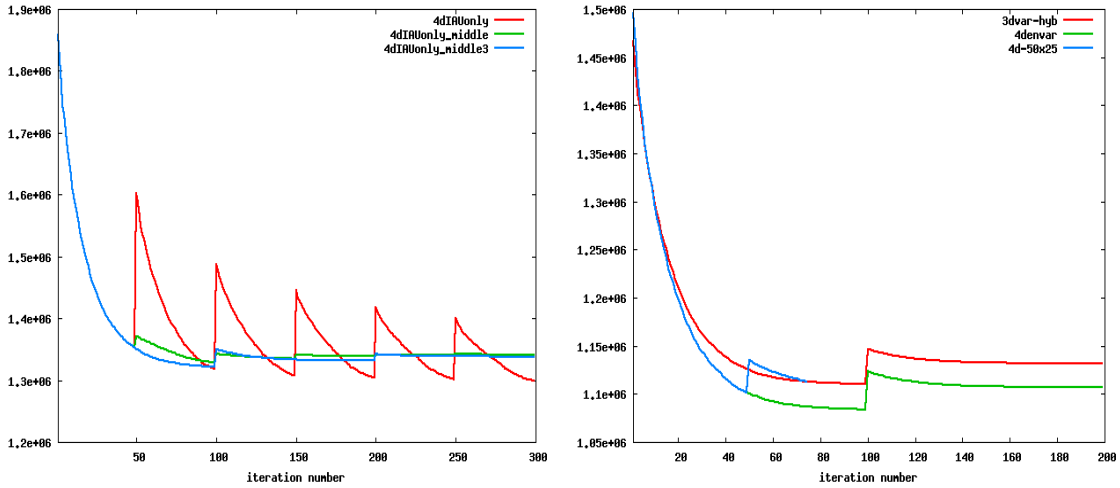


**Figure 10:** Evolution of the cost function for various configurations of middle- and outer-loops for Hybrid 4D-EnVar (left panel), and for configurations of single outer-loop with different middle-loop strategies to determine acceptable number of iterations in 4D giving that gives comparable convergence level to Hybrid 3D-EnVar avoiding possible overfitting (right panel). Results are for two different time periods, and shown for single cycle only, in both cases.

Regardless of the preconditioning procedure, having to integrate the nonlinear model at each outer iteration of a legitimate outer-loop strategy entails a computational cost not presently justifiable in GEOS ADAS. Under a middle-loop strategy there is still the question of how many inner iterations should be necessary for acceptable convergence, without an accelerated preconditioning. The right panel of Fig. 10 gives the rationale for the choice of iterations in the configuration of the GMAO Hybrid 4D-EnVar analysis. The figure compares the evaluation of the cost function as the minimization takes place in the following three cases: Hybrid 3D-Var (red curve), Hybrid 4D-EnVar (green), and Hybrid 4D-EnVar with a reduced set of iterations (blue curve). It is important to emphasize that in these experiments the guess fields are identical at the start of the minimization. That is, these experiments are all from the same, single, non-cycling case. Comparing Hybrid 3D-Var with Hybrid 4D-EnVar we see the typical overfitting of 4D strategies (cf. Trémolet 2007b). After a few trials it was found that setting the minimizations of the two inner-loops of the Hybrid 4D-EnVar configuration to 50 and 25 results in a convergence level similar to that obtained with Hybrid 3D-Var. This exercise has been repeated for different choices of guess fields and observation (i.e., time windows around different synoptic times) and all lead to roughly the same conclusion: two inner loops with 50 and 25 iterations, respectively, amount to a convergence level comparable to that of Hybrid 3D-Var.

The similarity in the convergence level of both Hybrid 3D-Var and Hybrid 4D-EnVar does not mean these two procedures generate equal analysis. If nothing else, recall that unlike in 3D approaches, a 4D approach provides high-frequency corrections to the background. Single observation experiments of non-cycled analyses aimed at illustrating the difference between these two approaches are plentiful in the literature. Here, we illustrate this difference using results from two fully integrated 3D and 4D assimilation

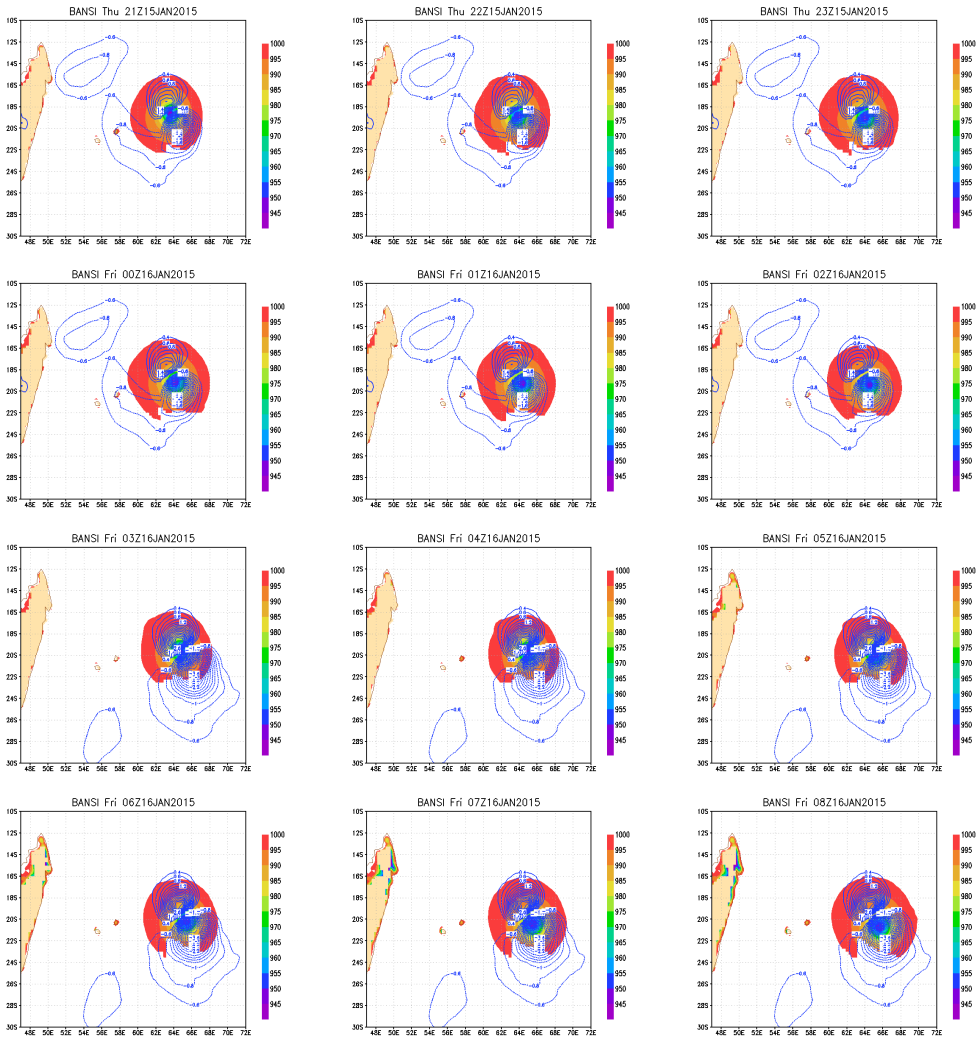


**Figure 11:** Hourly evolution of sea level pressure, and corresponding increments, from Hybrid 3D-Var of tropical storm Bansi off the east coast of Madagascar around mid January 2015. The shading shows sea-level pressure (hPa), and the contouring shows the corresponding sea-level pressure increment for the hybrid variational analysis. Snapshots are taken at hourly frequency, with date and time indicated in each panel.

experiments, when the whole observing system is present. Figures 11 and 12 show hourly snapshots of sea-level pressure (shaded) and increments of sea-level pressure (contours) from Hybrid 3D-Var (Fig. 11) and Hybrid 4D-EnVar (Fig. 12) for tropical storm Bansi off the east coast of Madagascar around mid-January 2015. A close look at the figures shows the sea-level pressure moving toward the west in each one of the consecutive snapshots. However, the increments of Hybrid 3D-Var are only seen to change (move) every 6 hours, and then remain the same for all hours within the next 6-hour interval, whereas the increments from Hybrid 4D-EnVar are noticeably changing every hour. At the transitions from one 6-hour cycle to the next, say, from 02 UCT 16 January to 0300 UTC 16 January there is a subtle change in the nature of the increment (in both the 3D and 4D cases). This is a consequence of the updated increment obtained from the solution of the (3D or 4D) assimilation problem over the new time window, which incorporates new incoming observations.


**Figure 12:** As in Fig. 11, but for Hybrid 4D-EnVar.

### 1.2.4 Evaluation of ensemble spread

The size of the ensemble necessary for the development of a reliable hybrid data assimilation system is still a matter of some debate in the NWP community. Although various studies examine the effects of an increase in their ensemble size an optimal size is far from being agreed upon. Indeed, the number of ensemble members varies rather wildly among different NWP DA systems. At the time of this writing, NCEP has 80 members in the EnSRF and Hybrid 4D-EnVar system[10]; the U. K. Met Office uses a 44-member ensemble in its global ensemble transform Kalman filter (ETKF); ECMWF uses 25 members in its ensemble of 4D-Var systems; and Environment Canada has 256 members in the EnKF supporting its Hybrid 4D-EnVar system (see Kleist *et al.* 2018, for the configuration of these and other operational systems). As indicated in Table 1, GMAO uses 32 members in its EnSRF, in both Hybrid 3D-Var and Hybrid 4D-EnVar options. A comprehensive comparison of its implied ensemble background error covariance, with say NCEP's, is beyond the scope of this manuscript. Still, it is worth mentioning that an effort has been made to get some sense of the reasonableness of the GMAO choice of ensemble size. It is not too complicated to get an idea of how the ensemble spread compares between GMAO and NCEP; comparing the correlations is somewhat more involving and not part of what we have attempted to do. Although the effects of size are more likely to

[10]This is the same number of members used when they were running Hybrid 3D-Var in their operational suite.

be apparent in raw correlations than in spreads, application of covariance localization should help attenuate differences, though it might be useful to compare correlations at some point.

Focusing on ensemble spread, there are at least two ways of looking at it: (i) in observation space; and directly (ii) in physical space. In observation space, the well established whiteness of the innovations result

$$< \mathbf{d}_{ob}\mathbf{d}_{ob}^T > \;\approx\; \mathbf{HBH}^T + \mathbf{R}, \tag{9}$$

connects the sequence of observation-minus-background (OMB) residuals, $\mathbf{d}_{ob}$, with a projection of the background and observation error covariances onto observation space (e.g., Kailath 1968 and Hollingsworth and Lönnberg 1989), as long as background and observation errors are truly uncorrelated. Separating the background and observation error covariances from the available OMB residuals, $\mathbf{d}_{ob}$, is not possible, though combining these with observation-minus-analysis (OMA) residuals, $\mathbf{d}_{oa}$, as proposed by Desroziers *et al.* (2005), allows for estimates to be inferred from

$$< \mathbf{d}_{ba}\mathbf{d}_{ob}^T > \;\approx\; \mathbf{HBH}^T, \tag{10a}$$

$$< \mathbf{d}_{oa}\mathbf{d}_{ob}^T > \;\approx\; \mathbf{R}. \tag{10b}$$

In these expressions, the observation residuals are defined as $\mathbf{d}_{oa,ob} = \mathbf{y}^o - \mathbf{h}(\mathbf{x}^{a,b})$, with $\mathbf{x}^a$ and $\mathbf{x}^b$, representing the analysis and background vectors, and $\mathbf{H}$ the linearized form of the observation operator $\mathbf{h}$. When the observation residuals come from a purely ensemble-base analysis procedure, such as the EnSRF, (10a) provides an estimate of $\mathbf{HB}_e\mathbf{H}^T$, that is, the implied ensemble background error covariance, projected onto the observations.

In a deterministic analysis system, such as a variational system — hybrid or not — only a single realization of the observation residuals is available. In these circumstances, it is typical to make the ergodic assumption and replace the ensemble mean operator on the LHS of (10) with a time-averaged operator. In an ensemble-based analysis system, this assumption is either no longer needed or can be combined with the time mean operation to augment the sample used in the averaging operation.

Figure 13 compares error variances derived from a 9-day sample of OMA and OMB residuals from the hybrid assimilation systems of GMAO and NCEP. These samples where collected when both centers were using Hybrid 3D-Var, with 32 and 80 members feeding into their respective EnSRFs. The figure shows a variety of standard deviations derived from the residuals of radiosonde zonal wind (left) and temperature (right) for GMAO (top) and NCEP (bottom). The black curves are for standard deviations from the single realization of OMB residuals for the corresponding hybrid (deterministic) ADAS; the red curves show the same quantity but derived from the residual time series of the corresponding ensemble mean observers. Almost by construction, these two time series are nearly identical within the context of a given system, for a given variable. That is, the GMAO standard deviation for radiosondes zonal wind (upper left panel) derived from the deterministic residual time series (black curve) is very close to that derived from the ensemble mean observer residual time series (red curve). This resemblance is not too surprising to find for reasonably well tuned systems using an analysis re-centering strategy (Sec.1.2.2). The same is noticeable in the results obtained from the NCEP observation residuals. A little more surprising, but encouraging, is the resemblance of the standard deviations between both GMAO and NCEP systems. It is important to emphasize that although GMAO and NCEP share the analyses software for GSI and EnSRF, the configuration of their corresponding DA systems is completely different, with differences being even larger when it comes to the corresponding AGCMs, their grid definition, hydrodynamical cores, physical packages, and more.

Of greater interest here is what can be obtained from the Desroziers *et al.* (2005) expression (10a) when examining the observation residuals of the ensemble data assimilation. An estimate for the background error standard deviation [square-root of the diagonal of the RHS of (10a)], together with the estimated observation error standard deviation following from (10b) appear in Fig. 13 for radiosonde residuals of zonal wind (left) and temperature (right) for both GMAO (top) and NCEP (bottom) ensemble assimilation systems. The blue curves are for the ensemble background error standard deviation, and the cyan curves are for the corresponding observation error standard deviations. There is again, striking similarity in the estimates

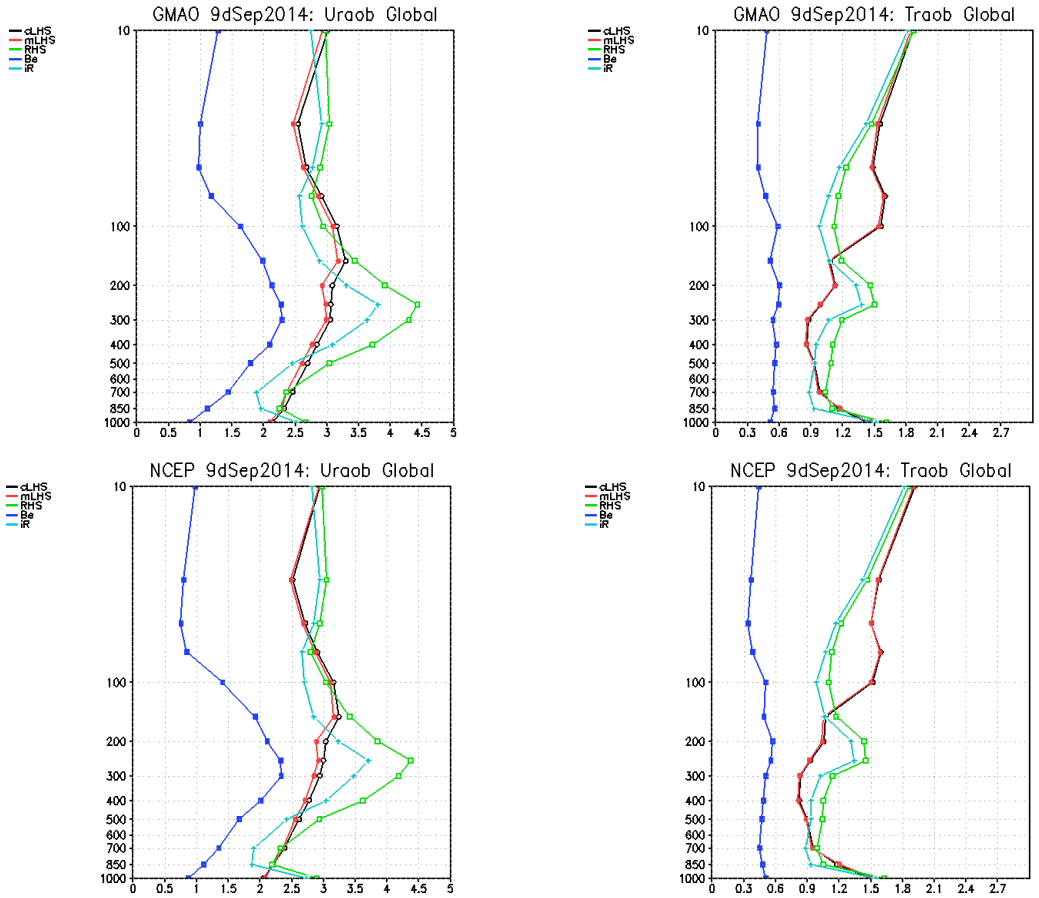


**Figure 13:** Observation-space ensemble spread estimation: First nine days in September 2014 of radiosonde residual statistic calculated from the GMAO (top) and NCEP (bottom) Hybrid 3D-Var systems. Curves are standard deviations of: cLHS (black curves) estimate of LHS of (9) calculated using OMB residuals from the central hybrid, mLHS (red curves) estimate of LHS of (9) calculated using OMB residuals from ensemble mean observers; Be (blue curves) estimate ensemble background errors from LHS side of (10a); iR (cyan curves) estimate of observation errors from LHS of (10b); the curves labeled RHS (green curves) correspond to the sum of both terms estimated from (10). Estimates Be, iR, and corresponding RHS are derived from times series of 32- and 80-member *ensemble* residuals from GMAO and NCEP, respectively.

between the two independent systems. GMAO and NCEP specify the same radiosonde observation errors and apply a similar error inflation approach to the radiosonde observations, and in this respect it is not too surprising to obtain similar observation error estimates from (10b). The similarity between the OMA and OMB residuals ensemble and the corresponding estimate of observation error standard deviation can only lead to similar estimates of ensemble background error standard deviation as seen by the blue curves in the figure. This seems to corroborate the fact that even with a 32 member ensemble, the GMAO system is able to simulate ensemble background error standard deviations with similar characteristics to those of the NCEP 80-member ensemble.

Examining the ensemble spread in physical-space is a much more direct exercise than estimating it through use of (10). Figure 14 shows globally-averaged ensemble spreads in zonal wind (left) and virtual temperature (right) as a function of pressure from both GMAO and NCEP systems (note difference in scales in all panels). These spreads are an average of the first 10 day periods: of January 2018 for NCEP global operational system and GMAO near-real-time (FP) system (green curves). A second case for GMAO is also shown (blue curves) covering the same September 2014 case considered in experiment used to derive the GMAO observation-space estimates displayed in Fig. 13. The shaded areas around the curves represent the one standard deviation variability in the spreads over the periods considered. When comparing GMAO and

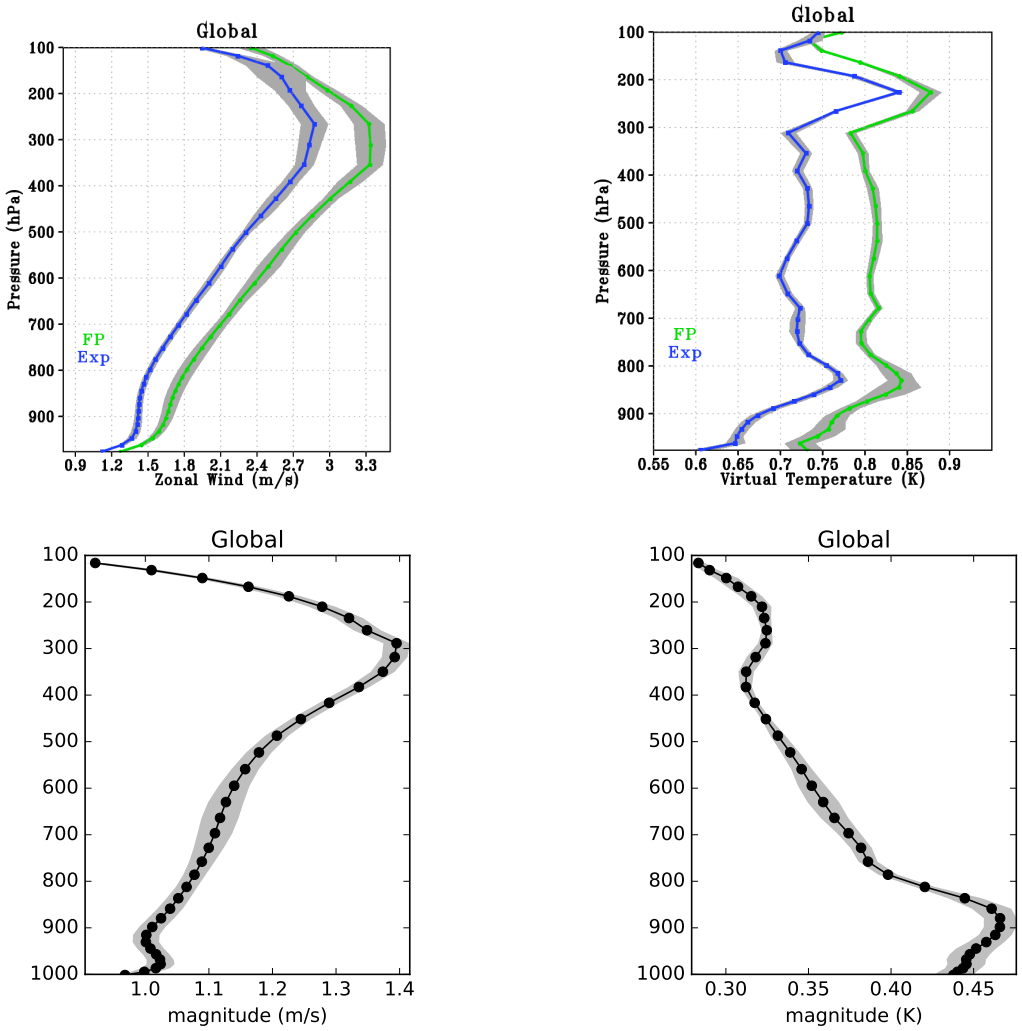


**Figure 14:** Physical-space ensemble spread: globally- and timely-averaged ensemble spreads for two cases from GMAO's system (top) and for NCEP's global operational system (bottom). Spreads are 10 day time averages: for NCEP, over initial part of January 2018; for GMAO, green curves (FP) are also over January 2018 for our FP system, and the blue curves (Exp) are for the period of September 2014 consistent with observation-space estimates shown in Fig. 13. Spreads are shown by the solid lines for zonal-wind (left column) and temperature (right column); the shaded areas represent the variability in the spreads over their respective 10 day periods. Results are plotted at corresponding model levels. Figures from NCEP are courtesy of Rahul Mahajan.

NCEP spreads we should recall how ensemble inflation is handled in these two systems. In the September 2014 cases used for the illustration in Fig. 13 both GMAO and NCEP were *only* applying the simple additive inflation mechanism based on NMC-like perturbations discussed earlier. In the January 2018 cases, the inflation schemes differ: NCEP has fully replaced its additive inflation procedure with a stochastic tendencies approach (see Palmer *et al.* 2009 and also Kleist *et al.* 2018); GMAO still uses the same additive inflation procedure[11]. With this in mind, the first noticeable difference in the physical-space spreads between GMAO and NCEP in Fig. 14 is their magnitude, with GMAO spreads being almost twice as large as NCEP's at certain levels. Still, much of the vertical structure in the spreads is rather comparable, with some differences observed near the boundary layer, which could well be explained by the difference in the

[11] Work is presently being done to have the AGCM ensemble members stochastically perturbed though in all likelihood we will end up with a blend of the two procedures; this will appear elsewhere.

corresponding ensemble inflation strategies, as well as in how these two systems resolve this layer. The variabilities in the spreads are not too dissimilar with, for example, temperature showing larger variability in the boundary layer in both systems and not much variability above that; and zonal wind showing about equal variability throughout most of the column in both systems, with GMAO having slightly greater variability at jet level. Comparing only the two GMAO cases in the top row of the figure we see an increase in spreads going from the September 2014 case (blue curves) to the current FP case (green curves). This is largely attributed to the difference in resolution of the ensemble; the test-case for the September period uses an ensemble at roughly 100 km (C90), whereas the January, FP case uses a 50 km ensemble (ref. to Table 1).

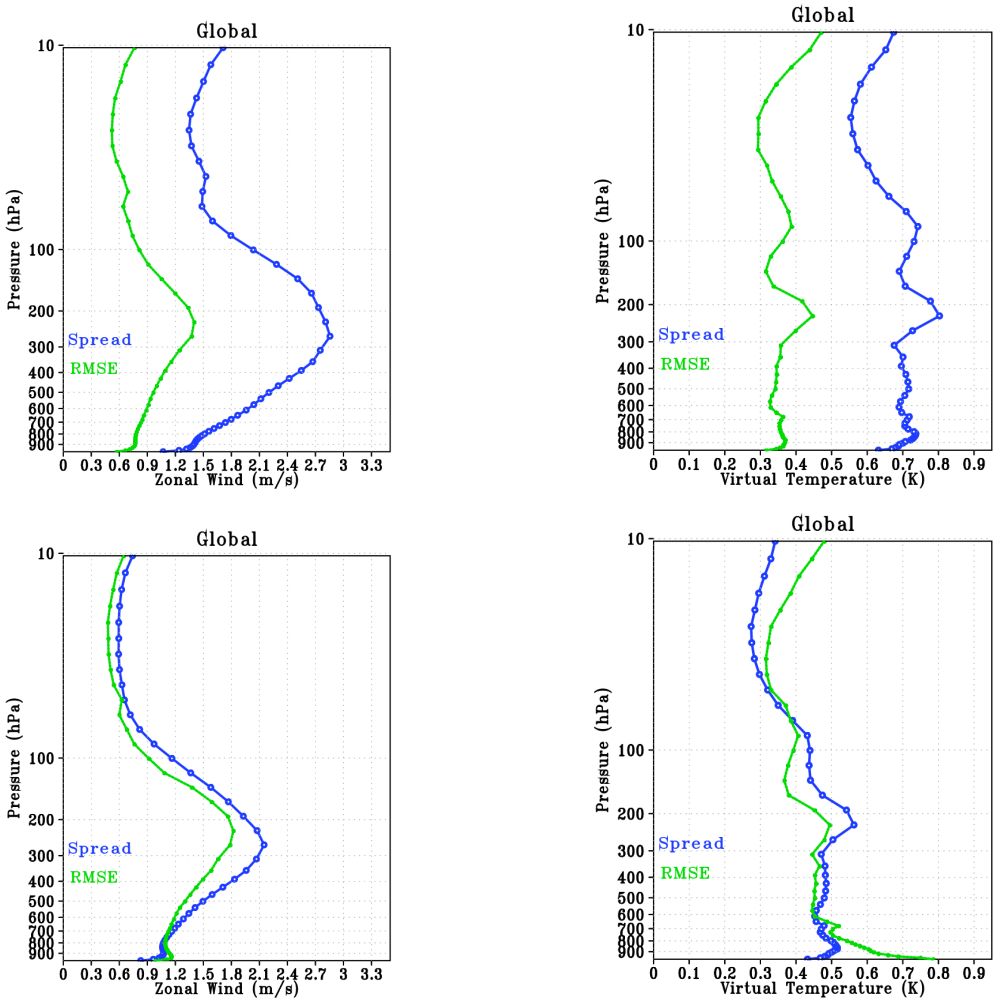


**Figure 15:** Two scenarios of ensemble adjustments are considered and physical-space ensemble spread is compared with RMSE of the ensemble mean in the 6-hour background: top row corresponds to default settings of experiment in Table 1; bottom row corresponds to a re-adjusted ensemble. Results for zonal-wind are on the left column; results for virtual temperature are on the right column. Blue curves in top row are similar to blue curves in top row of Fig. 14, except that here results are shown up to 10 hPa; notice difference in plotting scales.

Considering now only the September case in the top row panels of Fig. 14 we can compare the physical-space spreads with the corresponding observation-space estimates in the top row of Fig. 13 (blue curves of both figures). It is rather nice to see how the observation-space estimates corroborate reasonably well with the physical-space results — notice the difference in the plotting scales, also the fact that results in Fig. 13 include layers from 100 to 10 hPa, and that the observation-space results are shown at selective pressure levels whereas the physical-space results are shown at model levels up to 10 hPa. It is not surprising to see the physical-space results being slightly larger that the observation-space estimates since the latter are derived

over the rather sparse network of radiosondes. The vertical structures compare well, with the observation-space estimates even identifying a small reduction in temperature spread around 150 hPa.

Most hybrid data assimilation systems rely on known underdispersive ensembles. As one can imagine from the comparison between GMAO and NCEP spreads of Fig. 14, GMAO relies on an overdispersive ensemble. This can be better seen by comparing the ensemble spread with the root-mean-square error (RMSE) of the ensemble mean. Specifically, we follow Fortin *et al.* (2014) in comparing RMSE, $e$, with ensemble spread, $s$, in calculating the left- and right-hand-sides of

$$e \approx \sqrt{\left(\frac{M+1}{M}\right)\frac{1}{T}\sum_{t=1}^{T} s_t^2} \tag{11}$$

where the mean square error and the ensemble variance are given by

$$e^2 = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{I} w_i(\bar{x}_{i,t} - \bar{x}^v_{i,t})^2\,, \tag{12a}$$

$$s_t^2 = \frac{1}{(M-1)}\sum_{m=1}^{M}\sum_{i=1}^{I} w_i(x_{i,m,t} - \bar{x}_{i,t})^2\,, \tag{12b}$$

with the overbars standing for the ensemble mean; subscripts $i,m,t$ refer to the $i$-th grid-point, $m$-th ensemble member, and time $t$, respectively; $I$, $M$ and $T$, refer to the total number of grid points included in the spatial averaging, the total number of ensemble members, and the total number of time slots participating in the time averaging, respectively; the coefficients $w_i$ amount to grid weights such as the scaled cosines of latitudes; and $\bar{x}^v$ is a verification field, which in the illustration that follows is chosen to be the ensemble mean analysis.

Figure 15 compares the ensemble spread and RMSE of the ensemble mean for the 6-hour background fields of our 32-member ensemble. The blue curves display ensemble spreads, and the green curves display RMSE; results for zonal wind are in the left column, and for virtual temperature in the right column. The two rows of panels refer to two scenarios in the assimilation procedure. The top row corresponds to the default scenario of the experimental setting in Table 1, with spreads being analogous to what appears in the top row panels of Fig. 14, but now displayed up to 10 hPa. We see that in this case ensemble spread and RMSE are not lined up and show that indeed the ensemble is overdispersive. In the scenario of the bottom panel, the ensemble data assimilation has been adjusted such that now, both the spread and RMSE line up reasonably well. Earlier attempts in GEOS Hybrid ADAS to use such adjusted ensemble led to slight deterioration in the overall results of the hybrid system (not shown here); another attempt at a possible readjustment of the ensemble spreads is taking place as we write this document and a report of its consequences will appear elsewhere.

### 1.3 Evaluation of GEOS Hybrid 4D-EnVar

In what follows, we present a brief summary of results obtained from experiments performed during the test-phase before the upgrade to Hybrid 4D-EnVar. The illustrations here come from a small subset of these tests. Table 2 provides nomenclature and information on the experiments. Basically, two sets of experiments are considered over two separate time periods. The first time period covers the month of December 2015 during which two configurations of Hybrid 4D-EnVar are compared with Hybrid 3D-Var (considered as a control). The second time period covers the two months of October and November 2016 and compares a given configuration of Hybrid 4D-EnVar with 3D-Var (which is taken as the control in this case). We anticipate that experimentation with GEOS EVADAS echoes findings from other works in showing that upgrading from 3D-Var to Hybrid 3D-Var brings more noticeable improvement overall than when upgrading from Hybrid 3D-Var to Hybrid 4D-EnVar (e.g., Buehner *et al.* 2010a, and Buehner *et al.* 2010b). Indeed, works such as that of Clayton *et al.* 2013 find the best assimilation strategy to date to be Hybrid 4D-Var, which requires not only an ensemble of backgrounds but also the availability of tangent linear and adjoint models (or reliable simplified perturbation models) to participate in the inner loop minimization of the variational cost function

| Exp-Name | Strategy | Method | O-Loops | I-Loops | Resolution |
|---|---|---|---|---|---|
| December 2015 | | | | | |
| 3dHyb | Hyb-3dVar | IAU | 1 | 100x100 | As Exps Tab. 1 |
| 4dHyb | Hyb-4dEnVar | 4DIAU | 1 | 50x25 | As Exps Tab. 1 |
| 4dHyb-2L | Hyb-4dEnVar | 4DIAU | 2 | 1x50+1x25 | As Exps Tab. 1 |
| October-November 2016 | | | | | |
| 3dVar | 3dVar | IAU | 1 | 100x100 | As Exps Tab. 1 |
| 4dHyb | Hyb-4dEnVar | 4DIAU | 1 | 50x25 | As Exps Tab. 1 |
| From January 2017 onward | | | | | |
| GEOS.fp | Hyb-4dEnVar | 4DIAU | 1 | 50x25 | As NRT Tab. 1 |

Table 2: Experiments to evaluate Hybrid 4D-EnVar and decide on configuration to adopt for GMAO Forward Processing System. Experiments cover two time periods: December 2015 and October-November 2016. Control experiments for each period are highlighted in blue: Hybrid 3D-Var for December 2015; 3D-Var for October-November 2016. The final GEOS FP Hyb-4DEnVar configuration is shown in the last row. Detail on resolution of subcomponents of the system are laid out in Table 1.

(1). Although GMAO has working versions of these models (viz. its observation impact applications, e.g. Todling 2013; Holdaway *et al.* 2014), and it has cycled Hybrid 4D-Var, its associated computational cost is too high to support a near-real-time application such as GEOS FP. Considerable re-write of these codes and trajectory checkpoint re-tuning has taken place in the past several months (Holdaway, Errico and Kim, pers. comm.), where it now looks like we might be able to test a viable configuration of Hybrid 4D-Var for GMAO practical purposes. Here, we only report on Hybrid 4D-EnVar.

Both Hybrid 3D-Var and 3D-Var experiments in Tab. 2 use a single outer loop. In these, the minimization of (1) uses middle loop strategies with $2 \times 100$ inner iterations. In December 2015, two configurations of Hybrid 4D-EnVar are investigated: one named 4dHyb uses a single outer loop with a middle loop strategy with two outer iterations with 50 and 25 inner iterations each; and another experiment named 4dHyb-2L using two legitimate outer loops, each with inner loops with 50 and 25 iterations. In October-November 2016, the Hybrid 4D-EnVar system is configured as in the 4dHyb experiment of December 2015. All 3D experiments use the Bloom *et al.* (1996) version of IAU; all 4D experiments use 4DIAU, similar but not identical to Clayton *et al.* (2013) (details of the GMAO version of 4DIAU will appear in Takacs *et al.* (2018)). The last row in the table is for reference only and shows the final configuration of the Hybrid 4D-EnVar making up the GMAO FP system since early January 2017. Presently, all configurations of GEOS hybrid systems use a 50% split between the climatological and ensemble background error covariances in (4). Ensemble re-adjustment exercises such as the one briefly mentioned in the discussion associated with results in the bottom row of Fig. 15 suggest a potential benefit of re-setting the weights between these covariances to values likely to exceed a total of 100% (to appear elsewhere), along the same lines of the results of Bowler *et al.* (2017).

We start by looking at statistics of OMA and OMB residuals for the December 2015 experiments in Table 2. Figure 16 shows vertical profiles of monthly averaged radiosonde zonal wind (top) and temperature (bottom) OMA residuals over three regions, namely, global (left), tropics (center), and North America (right). The differences among the experiments are very subtle. When it comes to the OMA *biases* (dashed lines) the values are small and there really are no substantial Differences meriting much discussion. When it comes to the OMA *standard deviations* (solid lines), the 3dHyb experiment draws a little more closely to the observations than either of the 4D counterparts. This is not surprising given the choice of reduced number of iterations in the inner loop strategies of Hybrid 4D-Envar. As explained earlier, this is intentional and aims to avoid overfitting of observations by the 4D strategy. A similar comparison, but now for OMB,

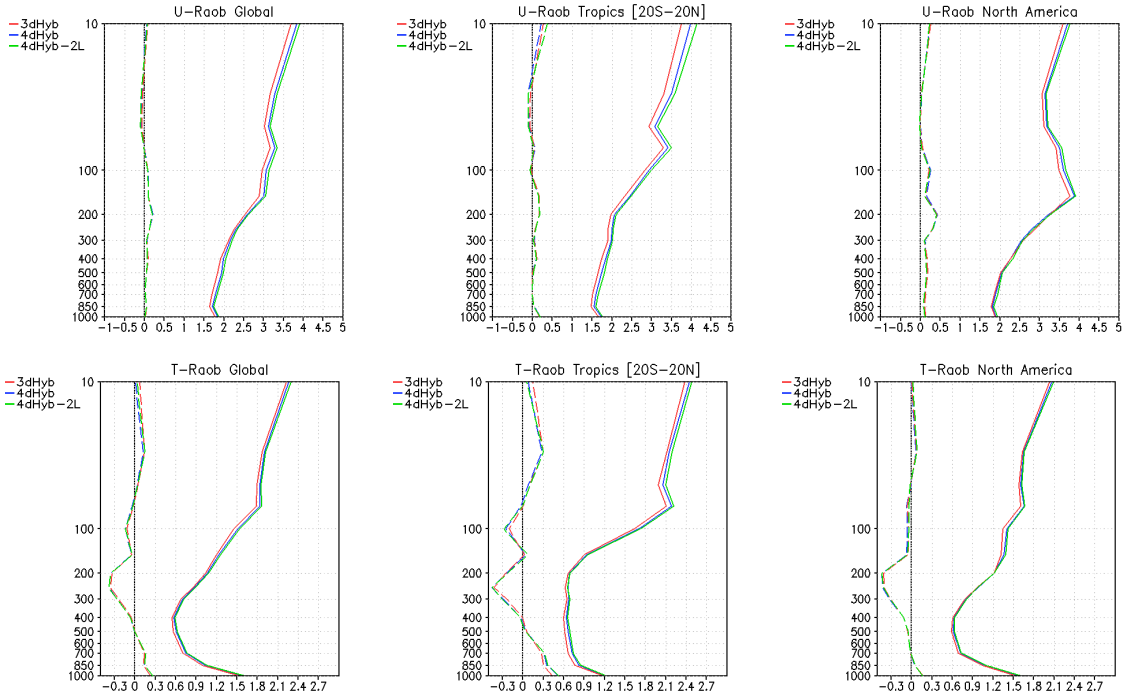


**Figure 16:** Regionally-averaged, December 2015 monthly mean radiosonde OMA residuals of zonal wind (top) and temperature (bottom) for experiments listed in Table 2. Regions shown are: Global (left), Tropics (middle), and North America (right). Dashed curves (floating around zero vertical line) are for OMA mean; solid curves are for standard deviations.

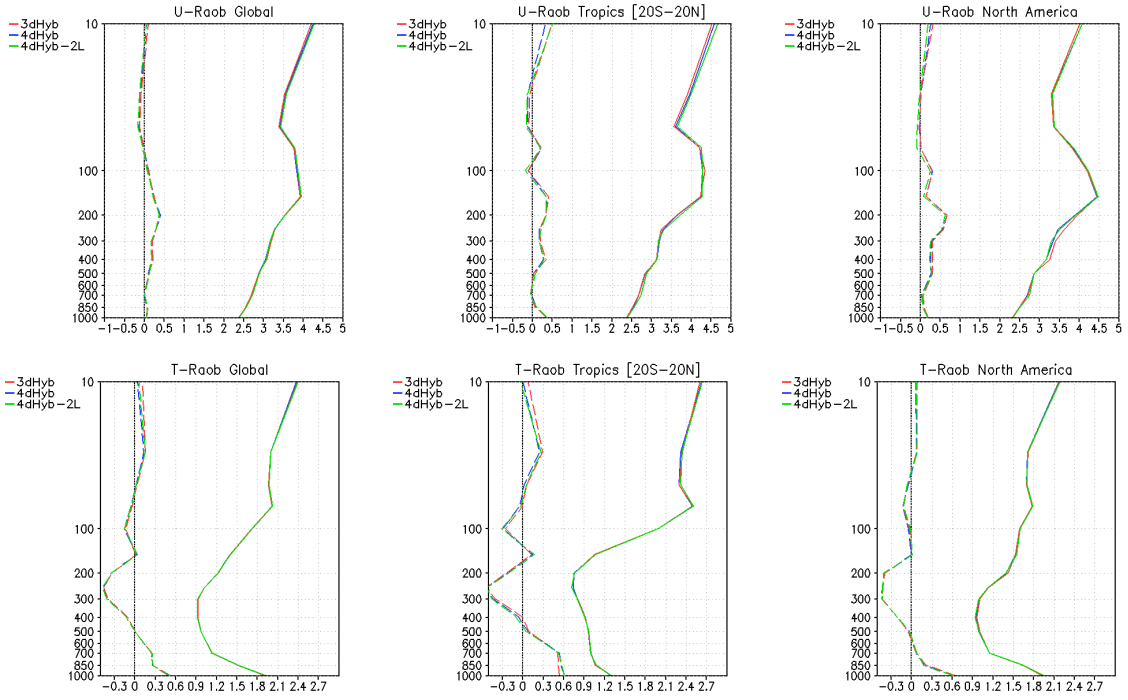


**Figure 17:** As in Fig. 16, but for OMB radiosonde residuals.

appears in Fig. 17. In this case, the biases in zonal wind are slightly increased from their OMA values; the biases in temperature are increased considerably around jet level, but no differences in either biases or standard deviations are accentuated or reduced when using Hybrid 4D-EnVar instead of Hybrid 3D-Var. We should say, in passing, that the OMB temperature biases identified here are a known issue in GEOS AGCM, a problem that is expected to be resolved in the upcoming replacement of its Chou and Suarez (1999) radiation package with the Rapid Radiative Transfer Model package (e.g., Clough *et al.* 2005 and reference therein).

It would seem from these results that not much improvement is obtained in going from Hybrid 3D-Var to Hybrid 4D-EnVar. However, examination of residuals from, say, aircraft observations reveals a clearer separation of results, with Hybrid 4D-EnVar bringing slight but definite improvements. Figure 18 shows bias and standard deviation for the December 2015 monthly averaged OMA (top) and OMB (bottom) zonal wind residuals from MDCARS and ASDAR aircraft, averaged over North America (left) and Europe (right), respectively. Though biases (dashed curves) are not very different among the experiments, the standard deviations (solid curves) clearly show Hybrid 3D-Var not to draw (OMA) as much to these observations, as well as not to predict (OMB) them as well as the 4D options. The reduced number of iterations in the Hybrid 4D-EnVar experiments does not prevent either the analysis or the model predictions from having improved fits to these observations over those of Hybrid 3D-Var. Another illustration of the improvement in aircraft residuals when going from Hybrid 3D-Var to Hybrid 4D-EnVar is shown in Fig. 20. This shows difference plots between the 3dVar and 4dHyb experiments of Table 2 in the gridded MDCARS (left) and ASDAR (right) zonal-wind standard deviations at 250 hPa. The calculations are such that the "bluer" the maps seem, the better Hybrid 4D-EnVar predicts aircraft observations as compared to Hybrid 3d-Var.

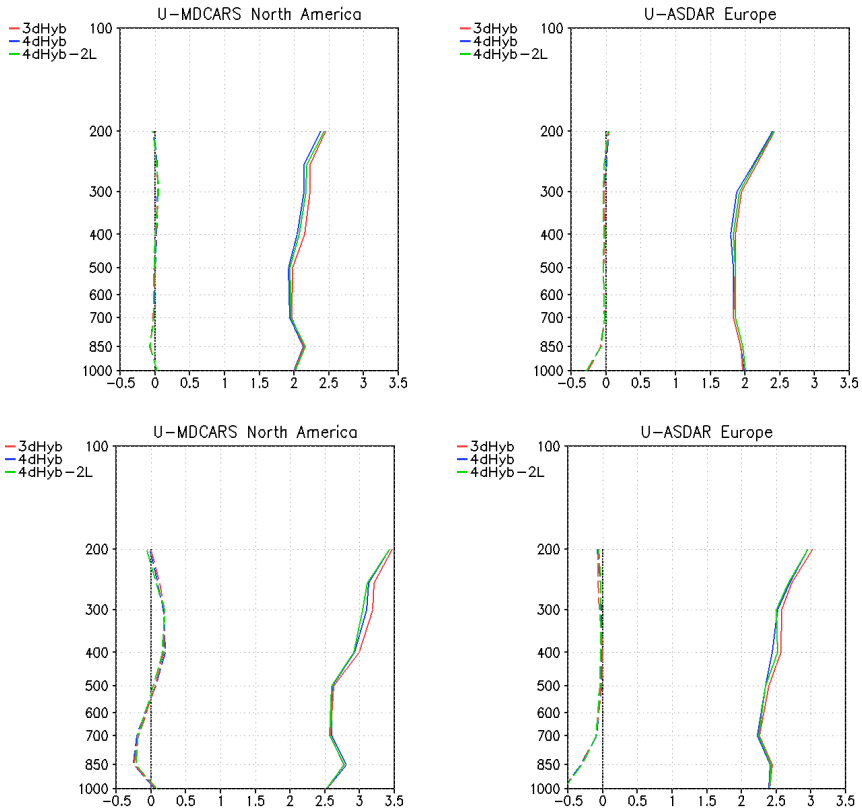


**Figure 18:** December 2015 monthly mean zonal winds OMA (top) and OMB (bottom) observation residuals from MDCARS aircrafts (left) over North America and ASDAR (right) over Europe: mean (dashed curves), standard deviation (solid curves) for experiments listed in Table 2.

As pointed out earlier, comparing Hybrid 3D-Var with Hybrid 4D-EnVar tends to show only small improvements as compared to experiments in upgrading 3D-Var to Hybrid 3D-Var. Evaluation of the experiments in Table 2 for October-November 2016 serves as corroboration of this point. Figure 19 shows regionally averaged OMA (top) and OMB (bottom) radiosonde residual statistics for the month of November 2016 for 3D-Var and Hybrid 4D-EnVar. In contrast to what we see in comparing Hybrid 3D-Var with Hybrid 4D-EnVar, now the benefits of hybridization show up as a reduction in the biases of both residuals (dashed curves). Examination of the standard deviations (solid curves) reveals Hybrid 4D-EnVar analyses to continue not to draw as hard to the observations as the analysis from 3D-Var. This similarity to what was found when comparing with Hybrid 3D-Var is simply a consequence of the choice made for the number of iterations in the inner loops of the minimization of the 4D problem. The improvement of Hybrid 4D-EnVar in the standard deviation of the OMB radiosonde residuals is considerably more evident here than when Hybrid 4D-EnVar is compared with Hybrid 3D-Var (compare bottom panels in Figs. 17 and 19).

It is also possible to examine the impact of observations on the analysis following Todling (2013). This amounts to an observation-space approach that uses the inverse of the observation error variances to define a measure for evaluating the contribution of various observing systems to the cycling assimilation. In its simplest form the so-called zero-hour impact evaluates the second term on the rhs of (1) at the initial and final steps of the minimization problem and then subtracts the latter from the former for each observation. The result is *typically* a negative quantity indicating the analysis *usually* reduces the error from the initial guess to the final solution with respect to each observation. The result can, for example, be grouped into the variables assimilated, such as wind, temperature, and radiance, or into the various observing instruments contributing to the analysis, or in many other ways. Fig. 21 displays time-averaged zero-hour impacts obtained over November 2016 for 3D-Var and Hybrid 4D-EnVar. Results are split into variable types (left) and instrument types (right). The most noticeable difference is in how both splits reveal that radiances have slightly larger impact in the 4D system than in the 3D system. This is expected of a system capable of better handling the evolution of errors in time (e.g, Rawlins *et al.* 2007). We also see the usual compensation effect at play in the impacts, with the 4D case displaying reduced impact of conventional observations as a reponse to the increased impact of radiance observations when compared with the 3D case.

A summary of the comparison between Hybrid 3D-Var with Hybrid 4D-EnVar for December 2015 and of the comparison between 3D-Var and Hybrid 4D-EnVar appear in Figs. 22 and 23, respectively. In the December case, since the two outer loop configuration is not quite suitable in practice at the moment, only results for the single outer loop study (4dHyb) are shown. The figures display percentage improvement (deterioration) of Hybrid 4D-EnVar over Hybrid 3D-Var and 3D-Var. All experiments are verified against observations (top panels), their own analyses (middle panels) and NCEP analyses (bottom panels), and the percentages in the figures are constructed from the relative comparison of Hybrid 4D-EnVar with the respective controls in each of the periods considered. Negative values indicate Hybrid 4D-EnVar to be an improvement over the respective 3D control experiments. A number of key score elements have been selected — typically including quantities of general relevance (such as 500 hPa geopotential height) and quantities our system has had trouble with in the past (such as the 250 hPa temperature, as indicated earlier). The scores are generated from 5-day forecasts for each of the assimilation experiments. Fig. 22 summarizes a full month of scores; Fig. 23 summarizes two months of scores. Statistical significance at 90% confidence level is shown by the thin cyan bars: blue and red bars larger in magnitude than the thin cyan bars correspond to statistically significant results. The results here again corroborate the anticipated fact that the hybrid system shows a larger improvement when compared against 3D-Var than with another hybrid flavor. Still, even when compared against Hybrid 3D-Var the 4D upgrade shows mainly statistically significant positive results. Without a doubt, the GMAO Hybrid 4D-EnVar represents a considerable improvement over its predecessor 3D-Var, across all three metrics considered in the figures. This is an encouraging and relevant result not only for our FP applications in support of NASA Instrument Teams, but also when it comes to support of reanalysis efforts (recall that both MERRA and MERRA-2 were based on traditional 3D-Var systems).

At present, the GMAO data assimilation system supports a variety of flavors of 3D and 4D assimilation

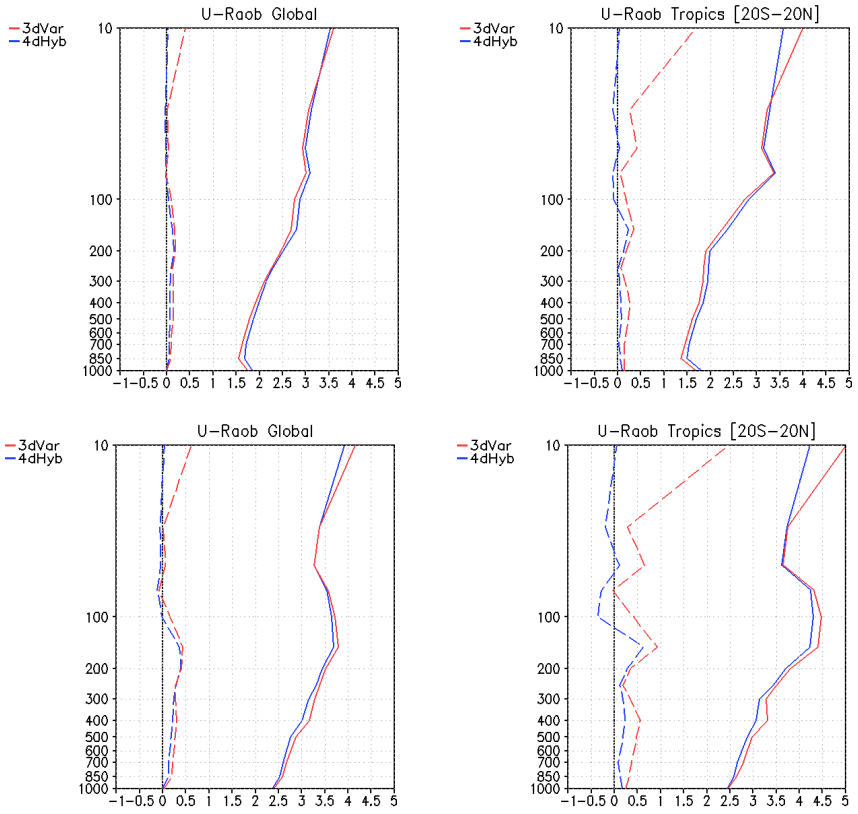


**Figure 19:** Regionally-averaged OMA (top) and OMB (bottom) radiosonde residuals, similar to those in Figs. 16 and 17, but now for the October-November period comparing 3D-Var with Hybrid 4D-EnVar as laid out in Table 2.

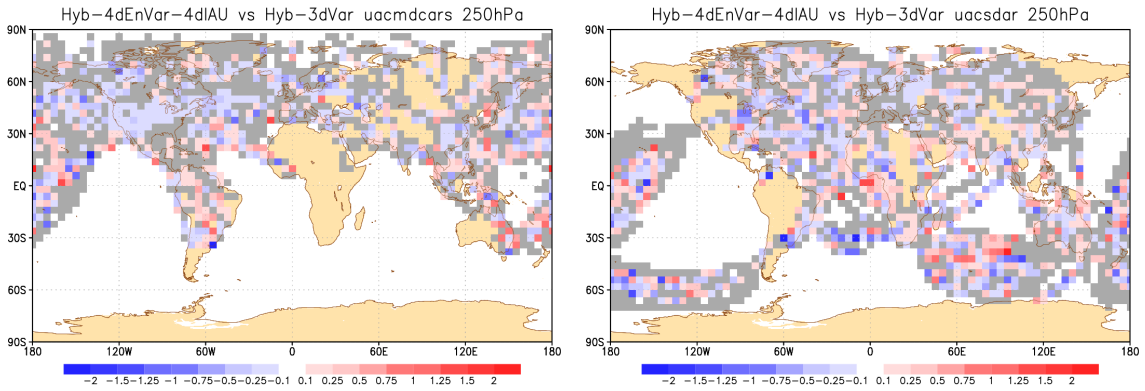


**Figure 20:** Differences between experiment (Hybrid 4D-EnVar) and control (Hybrid 3D-Var) of December 2015 monthly averaged standard deviations for MDCARS (left) and ACARS (right) zonal wind aircraft OMB residual statistics at 250 hPa; the red colors indicate control is closer to observations than experiment; the blue colors indicate experiment is closer to observations than control; neutral results are shaded grey.

procedures, from traditional 3D-Var, Hybrid 3D-Var, to 4D-Var, Hybrid 4D-Var, Hybrid 4D-EnVar, and others. Both MERRA Rienecker and coauthors (2011) and MERRA-2 Gelaro and coauthors (2017) are 3D-Var configurations of GEOS ADAS corresponding simply to the top grey-shaded block in Fig. 2; the present GMAO Forward Processing system comprises a Hybrid 4D-EnVar configuration running at a resolution

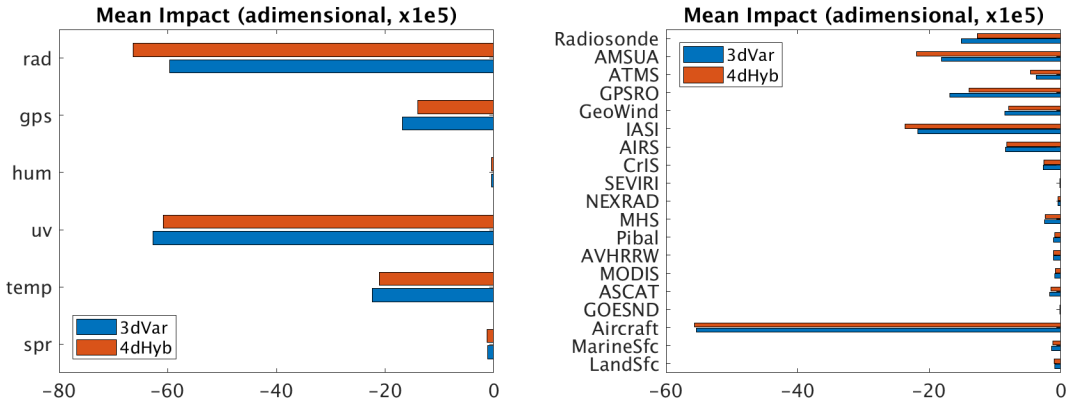


**Figure 21:** Zero-hour observation impact split in variable types (left) and instrument types (right) comparing 3D-Var and Hybrid 4D-EnVar (4dHyb experiment in Table 2). Results are monthly averages for November 2016.

considerably higher than that of either of these reanalyses — indeed, a single ensemble member of the FP system runs at the resolution of MERRA-2. The challenge of preparing our hybrid system to support the next GMAO reanalysis is non-trivial. The large changes in the observing system during the 30 or more years of reanalysis are bound to be reflected in the background error covariances represented by the ensemble and used as part of the background error covariance of the hybrid deterministic analysis. We expect to have to implement adaptive inflation and covariance localization procedures (e.g., Anderson 2009; Bishop and Hodyss 2011; Miyoshi 2011) to have improved representation of such background errors over the long time span of reanalysis. The requirement to have the next reanalysis incorporate an ocean-atmospheric coupled model poses even greater challenges. Here not only will we have to worry about the increase in cost of evolving each member of the ensemble as a coupled member, but we will need to investigate whether the number of ensemble members necessary for the hybrid covariance representation of the deterministic atmospheric analysis will suffice in aiding a fully ensemble-based (non-hybrid) weakly-coupled ocean analysis. In all likelihood some type of hybrid ocean analysis will have to be developed, perhaps as an expansion of the exisiting Ensemble Optimal Interpolation capability in the Ocean Data Assimilation System presently supporting our Seasonal Prediction System (Ham *et al.* 2013).

This concludes our brief description of the present state of science in this release of the GEOS Hybrid ADAS. As further progress continues to be made and new results become available this part of the document will be updated and new versions released. The remaining part of this document gives an overall idea of how GEOS ADAS implements its hybrid-variational strategy, giving special attention to how the atmospheric ensemble is created in each assimilation cycle. The document tries to serve as a User Guide providing details on the setting up of experiments, scripts, and the controlling environment variables.

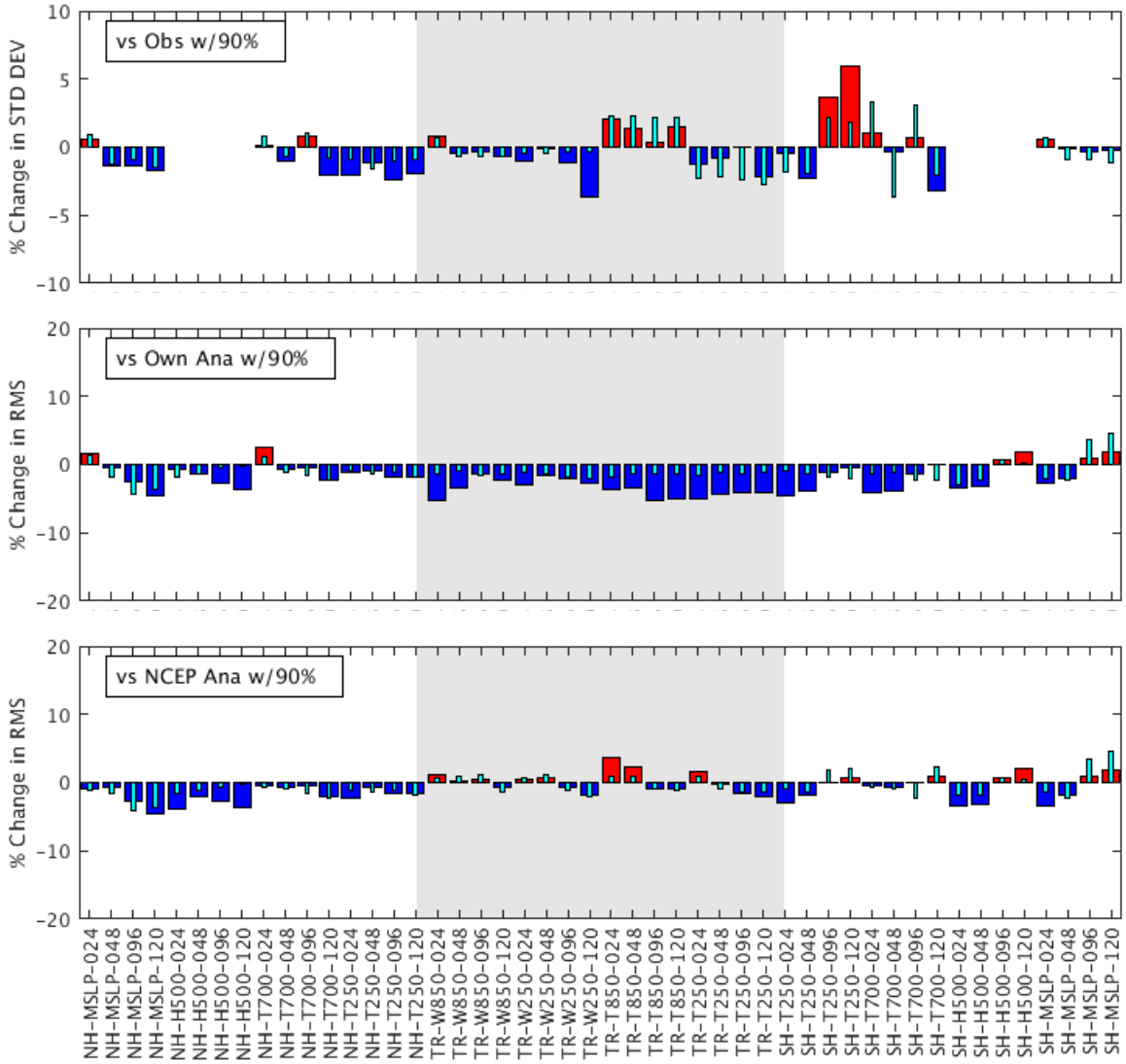


**Figure 22:** Percentage change in selective scores for when updating from Hybrid 3D-Var (control) to Hybrid 4D-EnVar (4dHyb experiment) during December 2015 as in Table 2. Scores cover the span of 5-day forecasts and use observations (top), own analyses (middle), and NCEP analyses for verification. Negative (positive) values, blue (red) bars, indicate improvement (deterioration) of results in experiment over control; thin (cyan) bars indicate 90% statistical significance in results (that is, results are statistically significant when thin bars are completely inside thicker bars).

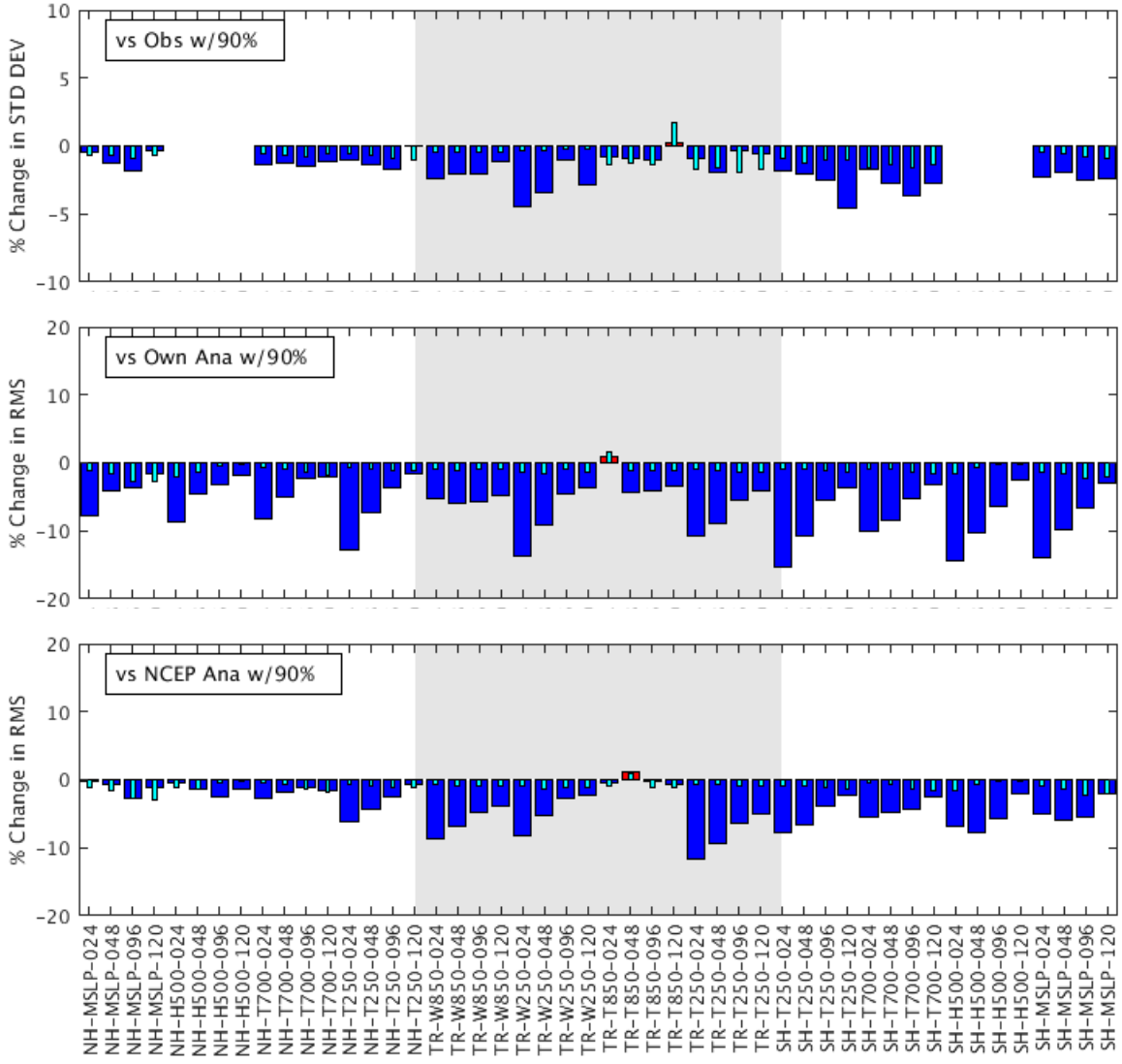


**Figure 23:** Similar to Fig. 22, but changes are now for when going from 3D-Var to Hybrid 4D-EnVar. Statistics in this case are collected over October-November 2016 with experiments configured as described in Table 2.

# 2 Overall Design

## 2.1 General description

This section presumes the reader has some familiarity with GEOS ADAS and its running mechanisms. As such, the following is not meant to be a comprehensive overview for how to run the GMAO data assimilation system. Readers are referred to the GMAO intranet website[12] for specifics related to non-hybrid (deterministic) ADAS.

The implementation of the ensemble ADAS capability within GEOS ADAS is rather non-intrusive in the sense that only minimal changes have been made to existing scripts and procedures already in place to support traditional 3D-Var. Specifically, only a minor set of changes have been made to the main job script driving GEOS ADAS, (`g5das.j`), and the so-called `analyzer`. The core of the driving scripts is controlled by `GEOSdas.csm`[13] and this too has suffered minute changes. The ensemble ADAS cycle is controlled independently by a job script named `atm_ens.j`. The procedures called by this script are shown in the flowchart of Fig. 24. A step-by-step description of these procedures is given later in this section.

As mentioned above, looking from the perspective of the central ADAS, the only difference between a *hybrid* analysis and a deterministic (non-ensemble) analysis is that the former needs to see the ensemble of background fields forming part of its background error covariance matrix. In practice, this corresponds to a very minor change to the `analyzer` that simply needs to know that a hybrid analysis is running and the location where to find the ensemble of backgrounds; the root of this information is set in the main ADAS driving script, `g5das.j`. The average user does not need to be aware of the details of the changes in `g5das.j`, since the procedure to setup GEOS ADAS experiments, namely `fvsetup`, automatically creates `g5das.j` and properly sets the required information depending on whether the user chooses to run traditional (non-ensemble) or hybrid ADAS experiments. In any case, for completion, we state here that three variables set in `g5das.j` control these options. The non-hybrid option has the following environment variables settings:

```
setenv HYBRIDGSI /dev/null
setenv STAGE4HYBGSI /dev/null
setenv RSTSTAGE4AENS /dev/null
```

and the hybrid option has the following settings:

```
setenv HYBRIDGSI $FVHOME/atmens
setenv STAGE4HYBGSI $HYBRIDGSI/central
setenv RSTSTAGE4AENS $HYBRIDGSI/RST
```

The variable `HYBRIDGSI` specifies the location of the ensemble members, the variable `STAGE4HYBGSI` specifies where to stage results from the central (deterministic) ADAS needed to run the ensemble ADAS, and the variable `RSTSTAGE4AENS` specifies where to place initial condition files from the central ADAS needed by the ensemble. The second and third variables can be collapsed into one. In the present settings of the hybrid system the only "initial condition" file from the central ADAS required by the ensemble is the file controlling the date and time of the cycle. The locations defined by these variables can be changed at will, not necessarily having to be defined as above; the hybrid default above is set to refer back to the location of the experiment itself, defined by the variable `$FVHOME`.

By default, traditional (non-hybrid) experiments set `g5das.j` to run one whole day of assimilation per batch submission, whereas hybrid experiments are, by default, set to run a single assimilation cycle per batch submission. Two environment variables found in the main driving script `g5das.j` control these settings:

```
setenv NSEGS 1
setenv NSTEP 1
```

[12] *https://gmao.gsfc.nasa.gov/intranet/personnel/rtodling/dasdev/GEOSDAS-UserGuide.htm; available upon request.*

[13] At the time of this writing, `GEOSdas.csm` is being split to allow GEOS ADAS to run under ecFlow (corresponding documentation soon to appear; see also https://software.ecmwf.int/wiki/display/ECFLOW/Tutorial).

This makes the scripts stop after one 6-hour cycle – this is the mode in which our near-real-time system runs. In hybrid mode, a new cycle can only begin after an ensemble of background fields is available. As discussed above, the `atm_ens.j` job script is responsible for the generation of these fields.

In its simpler mode of scheduling, the ensemble ADAS job is submitted at the end of the central ADAS job script and, accordingly, the central ADAS script is submitted at the end of the ensemble ADAS script. This mode of cycling the hybrid system does not exploit possible parallelism between the central and ensemble assimilation systems. In actuality, the ensemble ADAS does not need to wait for the whole central ADAS to finish, and vice-versa. The moments to synchronize the two systems are right before the central analysis begins, when one must have the required ensemble of backgrounds available; and right before the ensemble ADAS needs to re-center its member analyses about the hybrid GSI analysis which must then be available. Looking at Fig. 2, synchronization between the two systems must happen when the two red arrows hit the boxes they point to. Typically, both ensemble and central ADAS start from the same observation bias correction information which is updated at the end of every hybrid analysis. This mode of running is used in the GMAO Forward Processing System. The script controlling the synchronization of the various pieces is called `edas_scheduler.csh`; further details appear in Sec. 5.

Let us walk through the main steps in `atm_ens.j` by following the entries in the flowchart of Fig. 24. The first thing the ensemble job does is to look for the starting date of the cycle. This is done by consulting a copy of the date-time restart saved by the central ADAS under the location defined by the environment variable `$RSTSTAGE4AENS`. – see section 4 for specific configuration instructions. The script then goes on (almost) sequentially doing the following:

1. Generating perturbations for additive inflation.
2. Running the ensemble of observers.
3. Running the ensemble analysis.
4. Post-processing the ensemble of analyses.
5. Creating the ensemble of IAU-forcing terms.
6. Running the ensemble of initialized model forecasts.
7. Post-processing the ensemble of forecasts.
8. Triggering the central ADAS.
9. Archiving the ensemble ADAS output.

These steps make up the sequence of events in the default settings of the ensemble ADAS. Note that additional features available to the ensemble but not (yet) invoked in a typical cycle show up in Fig. 24 as double-dashed, marbled, boxes. These extra features are discussed later in Sec. 8. Section 9, on conventions, provides another view of how the ensemble ADAS is implemented. Specific details related to the steps laid out above are discussed in the remaining sub-sections of this section. Further information about each of the scripts encountered below is presented in the prologue of each script in the Appendix of this document.

## 2.2 Generating perturbations for additive inflation

Under the conventional mode of running the ensemble ADAS, it is necessary to obtain random perturbations to use as additive inflating factors applied to each ensemble analysis member. A database of NMC-like perturbations, 48-minus-24-hour forecast differences, has been generated covering a little over one year of forecasts from the so-called GEOS-5.7 series — a near-real-time version of GEOS that used to run 3D-Var

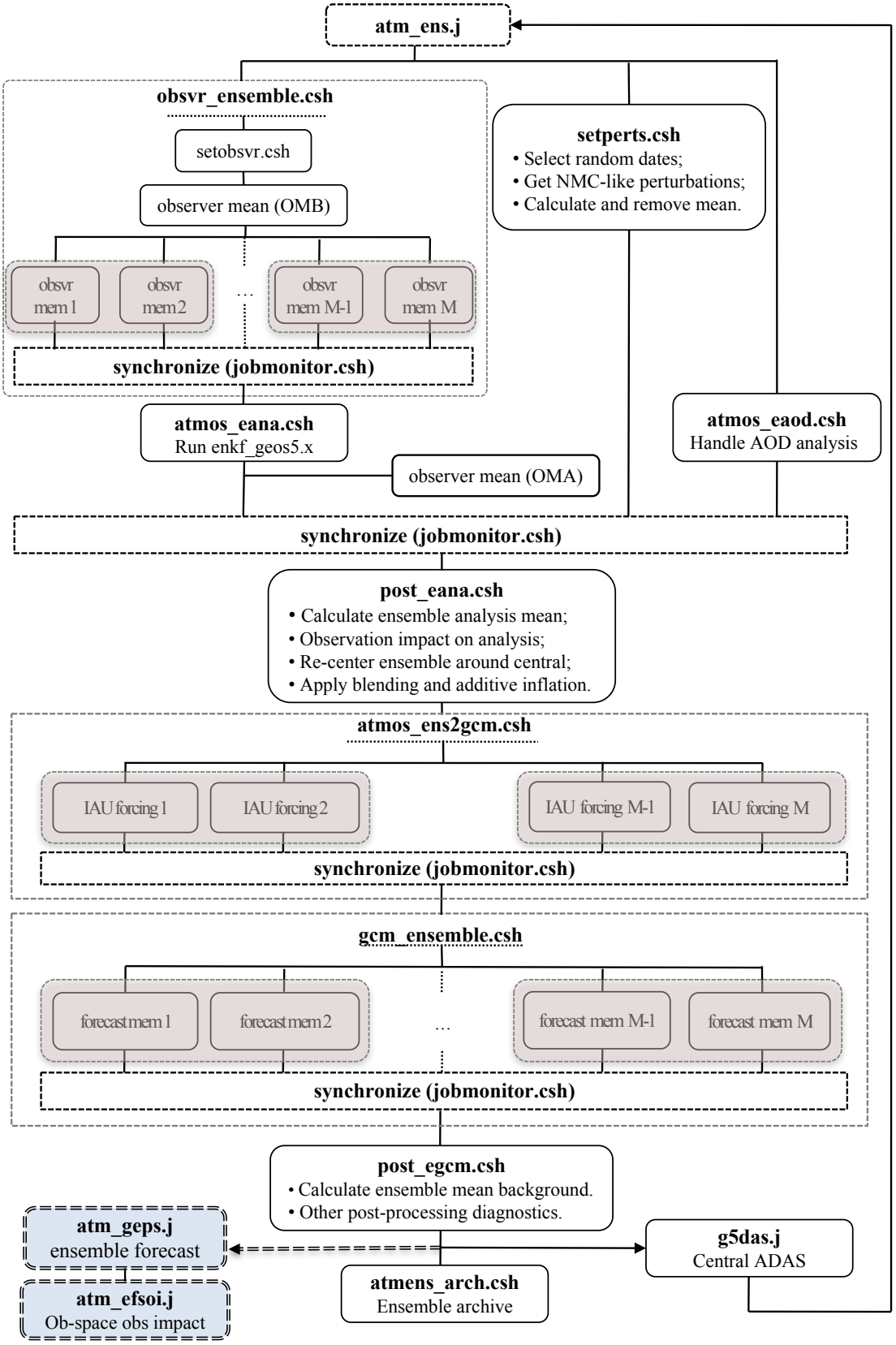


**Figure 24:** Flowchart showing the sequence of events in GEOS Ensemble ADAS (`atm_ens.j`) and its connection to the (central) hybrid ADAS (`g5das.j`). Double-dashed, marbled, boxes indicate alternative applications not normally called by default procedure.

with forecasts at a resolution of C360. The only regularly-available set of forecasts from GEOS-5.7 are those issued for the 0000 UTC cycles[14], therefore, only these are used as inflating perturbations[15].

As we will see in Sec. 4, most features of the ensemble ADAS are triggered by the presence of given resource files in the directory defined by the environment variable `ATMENSETC`, typically set to `$FVHOME/run/atmens`. Retrieving NMC-like perturbations from the database and preparing these perturbations for subsequent use is triggered in the main ensemble ADAS driver script by the following statements:

```
if ( -e $ATMENSETC/nmcperts.rc ) then
   if ( $DO_ATM_ENS || $RUN_PERTS ) then
      setperts.csh  ${EXPID} $nmem $anymd $anhms $TIMEINC $AENSADDINFLOC \
                    |& tee -a $FVWORK/setperts.log &
      if ($status) then
          echo "Main: failed in setperts.csh, aborting."
          exit(1)
      endif
   endif
endif
```

A check verifies the presence of the file `$ATMENSETC/nmcperts.rc`. This is a resource file with information related to the database of NMC-like perturbations: start and end dates of the database; location of the database; and whether or not seasonality is to be taken into account when retrieving perturbations randomly from the database (viz., NCEP operational hybrid system; J. S. Whitaker pers. comm.).

The c-shell script `setperts.csh` retrieves as many perturbations from the database as ensemble members being used. It then calculates the mean of the retrieved perturbations and generates a new set of perturbations with mean removed. This is done largely to avoid introducing biases in the current analysis since perturbations in the database are typically applied at a period likely unrelated to the synoptic time of interest. The controlling program doing the calculation of the mean is detailed in Sec. 7.3. Since these perturbations are needed only after the ensemble analysis completes, the work of fetching the perturbations and removing their mean can operate as a background job thus allowing the main script `atm_ens.j` to go on to its next task[16].

## 2.3 Running the ensemble of Observers

While the NMC-like perturbations are being de-biased, the main ensemble ADAS script moves on to run the ensemble of observers. Three steps are involved here: (i) retrieval of required inputs; (ii) execution of observer for mean backgrounds; and finally, (iii) execution of observers for each member of the ensemble. All these steps are handled by the script `obsvr_ensemble.csh`, which is invoked by `atm_ens.j` as shown below:

```
zeit_ci.x obsvr
obsvr_ensemble.csh $OBSCLASS $EXPID $anymd $anhms |& tee -a atm_ens.log
if( $status ) then
   echo "observer failed"
   exit(1)
endif
```

[14]Though some forecasts are available from 1200 UTC, they are not frequent enough to be conveniently used by the procedure that randomly selects perturbations.

[15]The current C720 4D-EnVar version GEOS issues regular 10-day forecasts from the 0000 UTC cycles, 5-day forecasts from the 1200 UTC cycles, and 30-hour forecasts for the 0600 and 1800 UTC cycles. These allow for possibly expanding the present database of NMC-like perturbations (see Sec. 12)

[16]This is the only procedure within the ensemble ADAS that is placed to run as background job; there are specific reasons for this. In general, we largely discourage processes to run as background jobs.

```
zeit_co.x obsvr
```

The required inputs for the observers are the observations, the background fields, and satellite bias estimation coefficients from a previous cycle. Retrieving observations from the archive is done in much the same way as when running the central ADAS: the `acquire_obs` is eventually called from within the setup script `setobsvr.csh` which in turn is called from within `obsvr_ensemble.csh`; `setobsvr.csh` also has the responsibility to obtain the satellite bias correction coefficients (see file `satbias.acq` usually placed under `$ATMENSETC`). The background files are obtained from the present location of the ensemble (specified through the environment variable `$ATMENSLOC`; see Sec. 4.1.1). Once all observations, auxiliary files, and resource files are available, the observer can be executed for the ensemble mean background[17]. This step simply entails running GSI with particular options for its parameter set, as defined in the resource file `obs1gsi_mean.rc`; no minimization is triggered in this case. The objective here is to calculate observation-minus-background (OMB) residuals for the mean (written out as typical "diag files") and to write out observations passing quality control. The selected observations are then taken in by each of the individual observer runs, for each member of the ensemble; the final step taking place within `obsvr_ensemble.csh`. The GSI parameter settings controlling this step are specified in the resource file `obs1gsi_member.rc`; here again, no minimization is triggered. By construction, all observers see exactly the same set of observations as used by the mean observer. In the end, each observer produces a set of so-called "diag files" with their corresponding OMB residuals. The member observers run parallel to one another; the level of parallelism is discussed later in this document.

## 2.4 Running the ensemble analysis

### 2.4.1 Atmospheric analysis: meteorology

With the mean and member OMB residuals available, the `atm_ens.j` script can now invoke the script controlling the atmospheric ensemble analysis. The corresponding statement in the driving script is as follows:

```
zeit_ci.x eana
atmos_eana.csh $EXPID $anymd $anhms |& tee -a atm_ens.log
if( $status) then
   echo "eana failed"
   exit(1)
endif
zeit_co.x eana
```

Depending on the settings (i.e., present resource files under `$ATMENSETC`) this procedure is capable of calling any one of a few different options of ensemble analysis schemes (see Sec. 4.1.4). In this introductory discussion, we assume the default settings are used and thus the atmospheric EnSRF analysis is called. The ensemble of OMB residuals is used to update the ensemble of backgrounds and create an ensemble of analyzed fields. Although the EnSRF is capable of updating backgrounds at flexible frequencies, the current implementation only updates the state at center-times of the assimilation window, these coinciding with the synoptic hours, corresponding to a 3D assimilation strategy.

### 2.4.2 Atmospheric analysis: aerosols

Since the Forward Processing System that preceeded MERRA-2, GMAO has implemented in its assimilation suite a three-hourly analysis of GOCART-derived aerosol optical depth (AOD; Colarco *et al.* 2010). The analysis of AOD is based on a 2D configuration of the Physical Space-space Statistical Analysis [PSAS; da

[17] This assumes the mean of the backgrounds to be available; see the post-GCM step below.

Silva, pers. comm.; see also, Cohn *et al.* 1998], and is referred to as the Goddard Aerosol Analysis System (GAAS). The 3-hourly update of the (current) 18 three-dimensional aerosol species of GOCART carried in the atmospheric model is done through the Local Displacement Ensemble (LDE) approach of Randles *et al.* (2017)[18]. Presently, only the deterministic, central, ADAS performs GAAS. This single realization of the three-hourly AOD analysis is used in conjunction with the LDE approach to update the 3D aerosol species in both the deterministic and ensemble model integrations.

The script controlling the ensemble of AOD analyses is named `atmos_eaod.csh`. In the framework just laid out, this script has the simple function of making the three-hourly AOD analyses from the central ADAS available to each of the members of the ensemble; in other words, presently, all members of the ensemble see the same AOD analysis. The resolution difference between the central ADAS and ensemble model integrations requires interpolation of the AOD fields, which takes place implicitly through GEOS MAPL utilities. The call of `atmos_eaod.csh` by the driving ensemble ADAS script looks as follows: (see also contents of file `aod4aens.acq`).

```
zeit_ci.x eaod
atmos_eaod.csh $EXPID $anymd $anhms 030000 |& tee -a atm_ens.log
if( $status) then
   echo "eaod failed"
   exit(1)
endif
zeit_co.x eaod
```

In this context, the acquire resource file `aod4aens.acq` specifies the location of the relevant AOD analysis files.

More generally, the script `atmos_eaod.csh` is also the entry-point to different flavors of possible generalization of the handling of AOD in the ensemble. One alternative to the present framework, already implemented as an option to this script, is to invoke the GAAS analysis component to analyze the background AOD of each ensemble member; the PSAS-based AOD analysis is rather low cost and presents no additional computational burden. Alternatively still is the possibility of enabling the EnSRF to perform the ensemble update of AOD. A preliminary implementation of this option (including the ability of the EnSRF to analyze AOD) is available as well (Buchard et al. per comm.). Upcoming versions of the hybrid system will introduce revisions to the AOD analysis in both the ensemble and central ADAS components.

## 2.5 Post-processing the ensemble of analyses

In the default configuration, the next step to take place in the flow of the driving script (Fig. 24) is the re-centering of the ensemble members around the central hybrid analysis, with subsequent application of additive inflation. This later step assumes that the NMC-like perturbations to serve as additive inflation terms are ready for use. Recall that these were being processed by `setperts.csh` that had been running as a background job while the ensemble of observers and the EnSRF had been running. At this point, there is need to synchronize the generation of the perturbations with the main driving script. This takes place by calling the `jobmonitor.csh` procedure in main:

```
set ah           =  `echo ${anhms} | cut -c1-2`
set ayyyymmddhh  =  ${anymd}${ah}
jobmonitor.csh 1 setperts.csh $FVWORK $ayyyymmddhh
```

The job-monitor script makes sure the (`setperts.csh`) script has finished successfully. The job-monitor script works like a barrier call in, for example, MPI programs, thus working to synchronize all running

[18]Note the LDE approach creates its own sample of members fully unrelated to the model-generated ensemble members of the EnADAS.

processes. This synchronization requirement is explicitly shown in Fig. 24. When all is complete, the post-analysis procedure can then be called:

```
zeit_ci.x post_eana
post_eana.csh $EXPID $anymd $anhms |& tee -a atm_ens.log
if( $status) then
   echo "post_eana failed"
   exit(1)
endif
zeit_co.x post_eana
```

Under default settings, the call above calculates the mean of the updated ensemble members. Complementary, more advanced settings are available that instruct the scripts to calculate observation-minus-analysis (OMA) residuals and, possibly, so-called 0-hour observation impacts based on the ensemble mean analysis following (Todling 2013; see Sec. 8.2). After the ensemble mean analysis is available, the post-processing continues on to re-center the analyzed ensemble members around the hybrid GSI analysis. Re-centering amounts to removal of the ensemble mean and addition of the central analysis to each member of the ensemble. During this procedure, additive inflation is also applied by scaling the de-biased NMC-like perturbations generated from `setperts.csh` and adding them to the re-centered analysis members. Section 7.2 discusses in greater detail what really happens under the covers of this step[19].

## 2.6 Creating the ensemble of IAU-forcing terms

As pointed out earlier, the ensemble analysis in the EnSRF is presently set to update the synoptic time backgrounds, which corresponds to a 3D-type assimilation strategy. In this context, a traditional implementation of IAU is used as the assimilation approach — this being the method by which the model sees the changes due to the analysis scheme. Therefore, when all member-analyses are available, the step to create the necessary AGCM input (restart) file carrying the IAU term is called by the main `atm_ens.j` script in the following way:

```
zeit_ci.x ens2gcm
atmos_ens2gcm.csh $EXPID $anymd $anhms |& tee -a atm_ens.log
if( $status) then
   echo "ens2gcm failed"
   exit(1)
endif
zeit_co.x ens2gcm
```

This procedure is responsible for creating the corresponding IAU forcing for each member of the ensemble. The script `atmos_ens2gcm.csh` is a wrapper controlling the program `mkiau.x`, which has the function of recognizing the underlying model grid (regular lat-lon or cubed) and converting analysis variables into variables that can easily be turned into model tendency terms. Furthermore, this program also incorporates a vertically integrated mass-wind divergence adjustment procedure, and possibly (but typically not exercised) the ability to apply a vertical damping to the increments. The options in this program are controlled in a resource file with the name `mkiau.rc.tmpl` present in the directory `$ATMENSETC` in the form of a template file[20]. Just as with the observer members, the member IAU-forcing terms are generated in parallel to one another; the details on the level of parallelism is discussed later on in this document.

[19] For example, when necessary, this step remaps the central analysis to the topography of each member; furthermore, this step applies vertical blending to maintain the stratosphere of the members as close as possible to that of the central GSI analysis.

[20] GEOS has a number of so-called templated resource files which define options to specific components of the system. Unlike resource files, which have a set of parameters that remain fixed through an experiment, "templated" resource files have a combination of pre-set parameters and options that can change from cycle-to-cycle or application call to application call. In the case of `mkiau.rc.tmpl` the increment valid date and time is a floating option.

## 2.7 Running the ensemble of initialized model forecasts

With the ensemble of IAU forcing terms available, an ensemble of AGCM integrations can be launched to perform the (initialized) assimilation of each of the members and create the background fields required for the following cycle of the ensemble ADAS and the following hybrid (central) analysis. The ensemble of model forecasts is controlled by the script `gcm_ensemble.csh`, called by main script as indicated here:

```
@ tfcst_hh   =  2 * $TIMEINC / 60
zeit_ci.x gcm_ens
gcm_ensemble.csh $EXPID $nymdb $nhmsb $tfcst_hh $ens_nlons $ens_nlats \
                 |& tee -a atm_ens.log
if( $status) then
   echo "gcm_ensemble failed"
   exit(1)
endif
zeit_co.x gcm_ens
```

By construction, the AGCM integrations cover the 6-hour IAU period corresponding to the assimilation window (defined by `$TIMEINC`) plus a 6-hour background generation period; the total 12-hour interval is specified in the call above by the variable `fcst_hh`. The 12-hour period controls the extent of the integration related to the background. Extending the ensemble of model integrations beyond 12 hours is controlled through the typical mechanism of resource file management, as done for the central ADAS (not by changing `$TIMEINC`; see Sec. 4.1.5).

With successful completion of the ensemble of AGCM integrations, a crucial step takes place in the main ensemble ADAS driver script at this point: the original (initial) ensemble is put to the side for archiving purposes; and the newly generated ensemble is moved out of the work directory and placed where the original (initial) ensemble was. The following statements in `atm_ens.j` perform this action:

```
/bin/mv $ATMENSLOC/atmens     $ATMENSLOC/atmens4arch.${nymdb}_${hhb}
/bin/mv $FVWORK/updated_ens   $ATMENSLOC/atmens
```

The important thing to realize is that this step avoids any copying of files; copying would take a considerable amount of time and be utterly inefficient — each member is associated with about twenty files among just backgrounds and AGCM restarts, not counting output from the analysis and other diagnostic output files possibly requested of each model integration.

## 2.8 Post-processing the ensemble of model forecasts

Once the ensemble of AGCM integrations is complete, everything needed to start the central ADAS is available. But, before the main ADAS script is launched, the ensemble ADAS post-processes the ensemble of background fields[21]. The AGCM post-processing step is called next:

```
zeit_ci.x post_egcm
post_egcm.csh $EXPID $nymdb $nhmsb $TIMEINC $FVHOME/atmens
zeit_co.x post_egcm
```

This entails calculating the ensemble mean of the background fields, as well as any offline diagnostics related to the ensemble, such as spread and RMS (see Secs. 7.3 and 7.4). Remember that the ensemble mean background is needed by the mean observer whenever the subsequent ensemble ADAS cycle begins again. Thus at a minimum, the ensemble mean background needs to be made available to the next cycle; this being calculated in the post-egcm setup above.

[21] In principle, there is no need to wait for this post-processing step to finish before launching the central ADAS and its corresponding hybrid GSI. However, this first release of the hybrid system does not take advantage of this level of parallelism due to some inefficiencies in the post-processing that will be tackled in future releases; see Sec. 12.

## 2.9 Triggering the central ADAS

The ensemble ADAS driving script has now reached the stage when the central ADAS job can be launched. The statements below show how this is handled in `atm_ens.j`:

```
cd $FVHOME/run
if( -e ${EXPID}_scheduler.j ) then
   touch $FVHOME/.DONE_MEM001_atm_ens.${yyyymmddhh}
else
   qsub g5das.j
endif
```

Notice there is a check associated with the scheduler, that is, when the ensemble and central ADAS are running in parallel all the ensemble driving script needs to do is indicate its successful completion — the batch submission of main components is controlled by the scheduler. In the sequential mode of submission, continuing on to the central ADAS simply amounts to submitting the central ADAS script, `g5das.j`, to the batch queue. Figure 24 illustrates the continuation of the cycle when the central ADAS job is launched.

## 2.10 Archiving the ensemble ADAS output

The only task to complete now is for the ensemble ADAS driving script to launch the archiving script call so the output of the ensemble ADAS can be saved in the mass-storage system. This happens in the final call of the `atm_ens.j` script:

```
atmens_arch.csh $EXPID $arch_nymd $arch_nhms \
                $FVHOME/run/atmens/atmens_storage.arc \
                eadas atmens4arch.${nymdb}_${hhb} \
               |& tee -a atm_ens_arch.${arch_nymd}_${arch_hh}z.log
```

As one can imagine, archiving output from the ensemble is non-trivial. Typical files can be somewhat large and the ensemble multiplies the total number of files to save rather dramatically. The present archiving procedure tries to be as efficient as possible, but improvements to this procedure are likely to be introduced with time. The archiving works by defining classes of files to be handled together and stacked in a single tar-file. Further details are found in Sec. 4.1.6. At a minimum, the ensemble of backgrounds, valid at the synoptic hour, should be saved. These allow for re-running the central GSI analysis (in a replay-like mode) without having to recreate the ensemble itself (see Sec. 8.3); these are also required by the adjoint-based observation impact machinery (see Sec. 8.1).

This completes the introductory description of the design and implementation of the GEOS Hybrid EVADAS. The sections that follow provide considerably more detail on each of the steps laid out above. Though we try to be as comprehensive as possible, readers must realize this system is a living entity: changes are frequently being made to expand capabilties and improve upon current mechanisms[22]

[22] Integration of the calls above in ecFlow is a change expected to replace `atm_ens.j` but not much of what this script controls; in time, a revised version of the controlling flow will be documented elsewhere.

This page intentionally left blank.

# 3 Repository Access and Installation Instructions

The following is a brief step-by-step set of instructions for how to access and compile the source code of GEOS Hybrid EVADAS[23].

**Checkout:** cvs co -r TAG MODULE, where TAG and MODULE define a particular version of interest. For example,

cvs co -r GEOSadas-5_17_0p5A GEOSadas-5_17

**Compiling:** After a fresh checkout, compilation can be accomplished by using the script `parallel_build.csh` residing under the GEOSadas/src directory. This script will prompt the user with simple questions; usually the defaults suffice.

**Central ADAS Setup:** After compilation completes, all relevant scripts and executables will be installed in the `bin` directory, usually under GEOSadas/Linux/bin (assuming a Linux machine architecture). The program `fvsetup` provides a way to configure a central ADAS experiment. This procedure prompts the user to multiple assimilation strategies options: 3D-Var, Hybrid 3D-Var, 4D-Var, Hybrid 4D-EnVar, and others.

**Ensemble ADAS Setup:** Selecting a hybrid option within the options of `fvsetup` requires the user to follow up with running the setup procedure associated with the ensemble ADAS. This is accomplished by invoking the script `setup_atmens.pl` found in the `bin` directory of the build. The following shows this setup usage:

```
NAME
     setup_atmens.pl - setup resources to allow running Hybrid ADAS

SYNOPSIS

     setup_atmens.pl [...options...] scheme
                                     expid
                                     aim
                                     ajm
                                     ogrid

DESCRIPTION


     The following parameters are required

     scheme   enkf or engsi
     expid    experiment name, e..g., u000_c72
     aim      number of x-grid points in Atmos GCM
     ajm      number of y-grid points in Atmos GCM
     ogrid    c, f, or C for low- or high-resolution Ocean GCM


OPTIONS

     -expdir       experiment location (default: /discover/nobackup/user)
     -atmens       location of ensemble members (default: FVHOME/atmens)
     -vtxrlc       use vortex tracker and relocator
     -h            prints this usage notice
```

[23] As of early 2018, GMAO has plans to move away from its CVS repository and migrate to Git. The instructions provided here are associated with the present official CVS repository.

```
EXAMPLE COMMAND LINE

     setup_atmens.pl enkf u000_C72 90 540 C

NECESSARY ENVIRONMENT

OPTIONAL ENVIRONMENT

AUTHOR

     Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GSFC/GMAO
     Last modified: 22Jun2016      by: R. Todling
```

# 4 Hybrid EVADAS Configuration

## 4.1 Ensemble ADAS

This section presents more specifics related to exercising the atmospheric ensemble data assimilation system that supports the GMAO hybrid ADAS (bottom grey-shaded box in Fig. 2).

### 4.1.1 Ensemble ADAS driving script

As we have seen, the driving script of GEOS ensemble ADAS is called `atm_ens.j`. This is the equivalent counterpart of the (hybrid) ADAS script `g5das.j`, but to control the atmospheric ensemble assimilation instead. Though there are similarities between these two scripts, their inner parts are substantially distinct. The `atm_ens.j` handles about ten steps from running the observer to archiving output from the ensemble ADAS cycle. All the steps are explicitly laid out – unlike in the ADAS (`g5das.j`) script which essentially calls the legacy `fvpsas` driver (or its alternative 4D-Var driver `g54var`), where all steps are hidden from the user.

As in the case of the central (hybrid) ADAS, once the batch script starts running, the work takes place in a temporary directory. However, unlike the central ADAS, the work directory for the ensemble is not a floating (TMPDIR-like) directory, randomly created each and every time the job script is submitted. The environment variable specifying the work directory is called `FVWORK`, just as in the central ADAS script, but in the case of `atm_ens.j` it is defined as follows:

```
setenv BIGNAME `echo "$EXPID" | tr -s '[:lower:]' '[:upper:]'`
setenv FVWORK /discover/nobackup/$user/enswork.$BIGNAME
```

The reason for fixing the work directory relates to the eventual need to re-submit the job script so it completes unfinished processes that might have failed the first time. The ensemble ADAS handles an incredibly large amount of work and processes. Sometimes, largely due to machine glitches, batch system issues, disc misbehavior, time-outs and more, the job may terminate unsuccessfully before the cycle ends. Fortunately, in the large majority of such cases, the design of the ensemble ADAS is such that it requires the user to simply re-submit the driving script `atm_ens.j` so that it picks up from where it left and completes the remaining tasks. At rare times, when things are really at odds with the computing environment, one cycle might stop more than once. Still, in these cases, users should simply verify that the reason for halting is indeed a glitch in the system, and re-submit the job once again. This ability to allow the driving script to pick up from where it left is the reason behind having a fixed (non-floating) work directory. Note that sometimes in this manuscript we refer to `FVWORK` as `ENSWORK` since this latter is the name of the environment variable used in the underlying scripts to refer to the ensemble ADAS work directory[24]

In a healthy termination, the fixed work directory is removed by the script. Therefore, one indicator of whether things have worked successfully or not is the absence or presence, respectively, of the work directory `enswork.$EXPID` in the user scratch ("nobackup") area after the `atm_ens.j` no longer shows up in the batch queue.

Developers of further and future features of the ensemble ADAS will find it useful to know that there are certain environment variables in the `atm_ens.j` job script that come in handy when working modifications into the system. These variables are defined near the top of the script `atm_ens.j` and are as follows:

```
setenv  DO_ATM_ENS    1
if ( $?eanaonly ) then
   if ( -e $FVHOME/run/${EXPID}_scheduler.j ) then
      setenv  DO_ATM_ENS    0
```

[24]Users should know that the locations "/discover/nobackup/" in the definition of `FVWORK` can be changed at will; also, users should be aware of the case *insensitive* definition of the work environment — as in the statements above.

```
   endif
endif
#    The following set specific pieces separately
#    Note: FVWORK better be defined by hand
#    ----------------------------------------------
if ( $?eanaonly ) then
   setenv  RUN_PERTS     1
   setenv  RUN_OBVSR     1
   setenv  RUN_EAANA     1
else
   setenv  RUN_PERTS     0
   setenv  RUN_OBVSR     0
   setenv  RUN_EAANA     0
endif
setenv  RUN_EAAOD      0
setenv  RUN_PEANA      0
setenv  RUN_ENS2GCM    0
setenv  RUN_AENSFCST   0
setenv  RUN_AENSVTRACK 0
setenv  RUN_ARCHATMENS 0
```

Their names are almost self explanatory. It is easy to imagine that with the above definition for the variable `DO_ATM_ENS`, the script will go over all the required ensemble ADAS steps (regardless of how the other environment variables are declared, right after `DO_ATM_ENS`). The variable check `$?eanaonly` is associated with when `atm_ens.j` is controlled by the scheduler; in sequential calls of the ensemble and central ADAS scripts this variable in not defined.

For example, if a user wants the job script to simply stop after completion of the ensemble of observers, one can set `DO_ATM_ENS` to zero and redefine the other variables as in:

```
setenv  DO_ATM_ENS    0
if ( $?eanaonly ) then
   if ( -e $FVHOME/run/${EXPID}_scheduler.j ) then
      setenv  DO_ATM_ENS    0
   endif
endif
#    The following set specific pieces separately
#    Note: FVWORK better be defined by hand
#    ----------------------------------------------
if ( $?eanaonly ) then
   setenv  RUN_PERTS     1
   setenv  RUN_OBVSR     1
   setenv  RUN_EAANA     1
else
   setenv  RUN_PERTS     1
   setenv  RUN_OBVSR     1
   setenv  RUN_EAANA     0
endif
setenv  RUN_EAAOD      0
setenv  RUN_PEANA      0
setenv  RUN_ENS2GCM    0
```

```
setenv  RUN_AENSFCST   0
setenv  RUN_AENSVTRACK 0
setenv  RUN_ARCHATMENS 0
```

This will force the driving script `atm_ens.j` to run only the steps controlled by the environment variables `RUN_PERTS` and `RUN_OBVSR`, associated with the generation of the NMC-like perturbations and the observers, respectively. In this case, after completion of the jobs associated with these two steps, the temporary work area remains available for the user to work within if necessary. Furthermore, the user can choose to run any number of steps as desired following this initial setting by simply activating environment variables associated with other procedures, e.g., running the ensemble analysis by setting `RUN_EAANA` to one, and re-submitting the driver script. As one can imagine, it would be a mistake to try activating, say, the variable `RUN_AENSFCST` which controls the model integrations before having gone through the steps prior to that.

### 4.1.2 Environment configuration controlling ensemble ADAS

Configuration of the ensemble ADAS takes place in the form of resource files and shell environment variables. This section covers the latter; specific resource-file configuration is treated separately in follow-up sections. All environment variables related to the ensemble ADAS are defined in the file `AtmEnsConfig.csh`, which is installed under the `etc` directory of the build. After the *ensemble* setup, a copy of this file resides under the directory $FVHOME/run/atmens. This file compartmentalizes environment variable definitions related to the various steps of the ensemble ADAS with general, globally applicable variables defined atop. A typical configuration of these variables is shown below.

```
setenv ATMENSLOC   $FVHOME            # locations of recycled files for
                                      #   atmospheric ensemble
setenv ATMENS4ARCH $FVHOME            # locations where dir with files to
                                      #   arch is held before arch
setenv ATMENSETC   $FVHOME/run/atmens # location of the ensemble-related
                                      #   resource files
setenv ATMENS_VERBOSE 1               # this can be put instead around each
                                      #   script call in atm_ens.j
setenv JOBMONITOR_MAXSLEEP_MIN 60     # maximum time (min) to wait for
                                      #   parallel job completion
setenv JOBGEN_NCPUS_PER_NODE 16       # general setting for jobgen script
setenv ENSPARALLEL 2                  # 0 - sequentially run ensemble
                                      # Otherwise ensemle runs in parallel mode
                                      # and in this case, the satellite bias
                                      # correction files can be treated in
                                      # either of the following two ways:
                                      # 1 - bias estimates from current
                                      #     hybrid cycle
                                      # 2 - bias estimates from previous
                                      #     hybrid cycle
setenv ATMENS_DO4DIAU 0               # 0=3DIAU; 1=4DIAU
```

The meaning of each of these environment settings is succintly explained in the comments declaration following the corresponding setting. It is important to know that the driving script sources this file after sourcing the file `FVDAS_Run_Config`, which defines general parameters for the ADAS. Therefore, the file `AtmEnsConfig.csh` is where overwriting of typical environment variables should reside. For example, the environment variable `OBSCLASS` is typically defined in `FVDAS_Run_Config`, which means that unless specified otherwise, the ensemble and the hybrid ADAS see the same observing classes (systems). There are

instances when the observing classes might differ between these two systems; a specific observation class choice for the ensemble should set OBSCLASS in AtmEnsConfig.csh (e.g., see Sec. 8.11).

Some of the variables shown above require a little further explanation. The location of the ensemble members, where the scripts expect to find the members of the ensemble, is defined by the variable ATMENSLOC. This is usually set to be $FVHOME/atmens. The location of the resource files related to the ensemble is separate from those of the regular ADAS, that is, the resource files associated with the latter are mainly found under $FVHOME/run, whereas those associated with the former are placed in the location defined by ATMENSETC, and typically set to $FVHOME/run/atmens, as indicated above[25].

The variable ENSPARALLEL defines the type of parallelism to be exploited by the ensemble ADAS scripts. A value of zero means that every task related to the ensemble ADAS runs sequentially; this has been useful during the initial phase of implementation of the ensemble ADAS, and it sometimes might help debug certain features, but it is only reasonable for a very small ensemble. Values different than zero imply that the ensemble internal tasks run in parallel. In this latter case, the variable ENSPARALLEL is used to simply distinguish between two options of handling the satellite (and possibly aircraft) bias correction files visible to the ensemble. The bias correction files are by default set to come from what the hybrid analysis derives: a value of 1 tells the scripts to expect this information to come from the current GSI hybrid analysis, and a value of 2 tells the script to expect the estimates to come from the previous analysis cycle. With the idea of having the ensemble observers and hybrid GSI observer "see" the same observations, the latter configuration is set as the default mode of running. Notice that only this default configuration is meaningful when the hybrid and ensemble ADAS are set to run parallel to each other and thus set to exploit the full level of parallelism existing between them (see discussion on scheduler, Sec. 5).

Lastly, the other variable to notice in the block of variables above is JOBMONITOR_MAXSLEEP_MIN. This sets the maximum time the job monitoring program allows any set of parallel tasks to run in batch. This is a critical variable, and its setting must be tuned to the behavior of the batch system, the size and resolution of the ensemble, and the level of parallelization of the main tasks. In the illustration setting above, the monitoring job allows one full hour for major processes to complete. For example, it means that all ensemble model integrations must end in less than one full wall-clock hour. This value is a high water-mark and can be adjusted at will, within reason. Section 7.6 discusses the intricacies of this variable more closely.

Beyond the basic settings above, the ensemble environment configuration file AtmEnsConfig.csh sets variables related to each of the tasks processed by the main driving script atm_ens.j as seen in Fig. 24. Following the sequence of events in the main driver, the next set of environment parameters found in AtmEnsConfig.csh relates to the retrieval and preparation of the ensemble of NMC-like perturbations necessary for the additive inflation procedure. In the example below these are set to run on the compute nodes of the NCCS (discover) batch queue using only a single CPU and a maximum wall-clock time of 1 hour (again, a wide water-mark).

```
# Environment setting to prepare NMC-like perturbations
# -----------------------------------------------------
setenv PERTS_QNAME compute
setenv PERTS_WALLCLOCK 1:00:00
setenv PERTS_NCPUS 1
```

Part of this procedure involves removing the mean of the NMC-like perturbations selected for any given analysis time; more on this appears in Sec. 7.3.

[25]The GMAO Operational Group, usually separates the root location of restarts (and other large files) from the root location of basic resource files and driving scripts; the former are usually kept in the scratch ("nobackup") area, and the latter are usually kept under a subdirectory of $HOME. A similar separation is allowed when running the hybrid variational-ensemble system.

Continuing down the sequence of events shown in Fig. 24, the next set of environment variables refers to running the observers. The settings below illustrate an acceptable configuration used when the ensemble of backgrounds is at 1-degree resolution. The wall-clock time (rather inflated), number of CPUS, and `mpirun`-related variables require no explanation; the variable `AENS_OBSVR_DSTJOB` is optional and requires some attention. After running the mean observer as an individual (32-CPU) batch job, when it comes time for the observer driving script (`obsvr_ensemble.csh`) to work on the individual member observers, an undefined (or absent) variable `AENS_OBSVR_DSTJOB` results in submission of as many 32-CPU batch jobs as there are members in the ensemble. On the other hand, when this variable is defined, as in the example below, the total number of batch jobs is determined as the total number of members divided by `AENS_OBSVR_DSTJOB`. For a 32-member ensemble the number of batch jobs is $32/4 = 8$, and since each observer requires 32 CPUs, each of the 8 batch jobs require a total of $32 \times 4 = 128$ CPUs.

```
# Environment setting to run the GSI observers
# --------------------------------------------
setenv OBSVR_WALLCLOCK 0:45:00
setenv ENSGSI_NCPUS 32
setenv MPIRUN_ENSANA  "mpiexec_mpt -np $ENSGSI_NCPUS GSIsa.x"
setenv AENS_OBSVR_DSTJOB 4
```

Notice that the actual observer work is performed by calling the GSI executable, `GSIsa.x`, for each member of the ensemble. The partitioning of the work in separate batch jobs is handled by the script `job_distributor.csh` installed in the `bin` directory of the build. Monitoring of the completion of the batch jobs associated with these tasks is done by the script `job_monitor.csh`, also installed in the bin directory of the build (more on this to appear later).

At this stage, following the flowchart in Fig. 24, both the ensemble of meteorological and of AOD analyses can be concurrently executed. As briefly explained earlier, presently, the AOD analysis is not actually performed for the members of the ensemble. Instead, the central, PSAS-based, AOD analysis is made available to each member of the ensemble. The settings in `AtmEnsConfig.csh` controlling the AOD analysis are as follows:

```
# ensemble GAAS and AERO EnKF
# ---------------------------
setenv NCPUS_AOD   8
setenv MPIRUN_AOD  "mpiexec_mpt "
setenv AENS_GAAS_OPT   1  # 1 members use central GAAS
                          # 2 analyze each member with PSAS
                          # 3 do (2), add EnKF-based AOD
                          #     analysis (off aod.or.concentrations)
                          # 4 EnKF-based AOD analysis
                          #     (off aod.or.concentrations)
setenv ATMENKFAERO_QNAME compute
setenv ATMENKFAERO_WALLCLOCK 1:00:00
setenv AENKFAERO_NCPUS 32
setenv MPIRUN_ATMENKFAERO "mpiexec_mpt -np $AENKFAERO_NCPUS enkf_aero.x"
```

Here, the main parameter to notice is `AENS_GAAS_OPT` which by default is set to use the central aerosol analysis. The alternative options available for comparison and testing are being further developed in coordination with Virginie Buchard and Arlindo da Silva (per. comm.); a version of the EnSRF enabled to handle the AOD analysis is available in the GMAO system (viz. executable named `enkf_aero.x`).

The meteorological analysis of the members is set to run the EnSRF by default. The program responsible for updating the members of the ensemble is named `enkf_geos5.x`. Only a single batch job is associated with this work; the illustration below shows the settings for running the 1-degree resolution case:

```
# Environment setting to run the EnKF analysis
# --------------------------------------------
setenv ATMENKF_QNAME compute
setenv ATMENKF_WALLCLOCK 0:30:00
setenv AENKF_NCPUS 96
setenv MPIRUN_ATMENKF "mpirun -np $AENKF_NCPUS enkf_geos5.x"
```

Once the ensemble (of meteorological) analysis completes, the ensemble needs to be re-centered and inflated. The environment variables controlling this part are shown below:

```
  Environment setting to re-center and inflate analyses
  -----------------------------------------------------
setenv AENS_ADDINFLATION 1         # apply additive inflation to members
setenv AENSADDINFLOC addperts      # location for additive perturbations
                                   #    (path relative to FVWORK)
setenv ADDINF_FACTOR 0.35          # additive inflation parameter
setenv RECENTER_QNAME general
setenv RECENTER_WALLCLOCK 1:00:00
setenv ENSRECENTER_NCPUS 4
setenv AENS_RECENTER_DST 4
```

The first variable, `AENS_ADDINFLATION`, tells the ensemble ADAS that inflation is to be applied to the analyses. As we learned earlier, the default is to use NMC-like perturbations for that. The location, inside the work area, where these perturbations are found is specified by the environment variable `AENSADDINFLOC`[26]. The factor used to scale the perturbations while adding them to the member analyses is specified by `ADDINF_FACTOR`[27]. The other variables above are self-explanatory given their similarity with variables already discussed.

The generation of IAU "restarts", containing the model-converted increments used to form corresponding analysis tendencies, is done by running as many instances of the program `mkiau.x` as members of the ensemble. Just as with some of the steps above, this can be done either by leaving out the environment variable `AENS_IAU_DSTJOB` from the list of defined variables and thereby submitting as many batch jobs as members, or by setting the variable `AENS_IAU_DSTJOB` as in the example below, which submits 8 batch jobs requesting $24 \times 4 = 96$ CPUs each to handle the calculation:

```
  Environment setting to create IAU forcing terms from each member
  -----------------------------------------------------------------
setenv AENS_IAU_DSTJOB 4
setenv IAU_WALLCLOCK 0:10:00
setenv ENSIAU_NCPUS 24
setenv MPIRUN_ENSIAU    "mpirun -np $ENSIAU_NCPUS mkiau.x"
```

[26] In retrospect, this variable should have been hidden from the user. This will be revisited in future releases.

[27] Eventually, this variable will either be moved to a resource file or disappear completely when considering possible adaptive procedures to inflate the members.

Running the ensemble of AGCMs is controlled by environment variables similar to those above. In the example below we chose, again, to have 8 batch jobs each controlling the execution of 4 simultaneous AGCM runs, within a job script requiring a total of $4 \times 48 = 192$ CPUS.

```
 Environment setting to run the ensemble of AGCM
 -----------------------------------------------
setenv AENS_GCM_DSTJOB 4
setenv AGCM_WALLCLOCK 1:00:00
setenv ENSGCM_NCPUS 48
setenv MPIRUN_ENSGCM  "mpirun -np $ENSGCM_NCPUS GEOSgcm.x"
```

The next step in the sequence shown in Fig. 24 is the post-processing of the ensemble of AGCM output. The following are the environment variables controlling this process:

```
# post-egcm calculations
# ----------------------
setenv PEGCM_NCPUS 4
setenv PEGCM_WALLCLOCK 1:00:00
setenv PEGCM_QNAME compute
```

The parameter `PEGCM_NCPUS` is used to distribute the calculation of means and statistics derived from the ensemble integrations. As one can imagine, many of these calculations can be done in parallel, particularly since they involve calculating statistics of snapshots of model output at various times; the processing of each time output can be handled independently of the others, as can the various output streams required of the model integrations (see Sec. 4.1.5).

Finally, the archiving procedure is controlled by the options below:

```
  Environment setting to archive results from an ensemble cycle
  -------------------------------------------------------------
setenv ENSARCH_FIELDS "eana,ebkg,ecbkg,erst,stat"
setenv ENSARCH_WALLCLOCK 2:00:00
setenv ARCHLOC /archive/u/$user
```

At the moment, the archiving mechanism is rather (NCCS-) discover-centric. Jobs handled by this procedure are automatically submitted to the *datamove* batch queue – specified in `jobgen.pl`. More importantly, however, is the method of defining collections of files to be archived. The example above refers to five minimal collections typically stored if users want to have the ability to (after-the-fact) reproduce any of the cycles of the ensemble: *eana* refers to the ensemble of analyses; *ebkg* refers to the ensemble of backgrounds (required by the hybrid GSI); `ecbkg` refers to the ensemble of chemistry backgrounds, required by the ensemble observer; `erst` refers to the ensemble of model initial conditions, required by the ensemble of AGCMs; and finally, *stat* refers to the mean, spread, and other statistics derived from the ensemble. More information about the archiving is available in the next section.

### 4.1.3 Configuration of the ensemble ADAS

After setting up a hybrid experiment (see Sec. 3), all resource files related to the atmospheric ensemble will be found under: `$FVHOME/run/atmens`. In principle, to run experiments blindly, the user does not need to know about the details discussed here. However, users conducting certain types of research and development will find the information below helpful when deciding how to exercise various features

and how to possibly expand existing ones. The current list of resource files associated with the default configuration of the ensemble ADAS is as follows:

**AGCM.rc.tmpl** - usual AGCM resource file, but controlling ensemble model integrations

**aod4aens.acq** - location/template-name of central aerosol analysis files

**AtmEnsConfig.csh** - sets all relevant Env Vars for the atmos-ensemble scripts

**atmens_storage.arc** - sets template names for archiving ensemble output

**atmos_enkf.nml.tmpl** - file with namelists related to EnSRF

**CAP.rc.tmpl** - usual AGCM resource file, but controlling ensemble model integrations

**GAAS_GridComp.rc** - controls use of PSAS-based AOD analysis for ensemble (off by default)

**GEOS_ChemGridComp.rc** - required if configuration of GCM-chem component different from central

**GSI_GridComp_ensfinal.rc.tmpl** - GSI-resource, controlling final ensemble of observers (for OMA)

**GSI_GridComp.rc.tmpl** - GSI-resource, controlling ensemble of observers (OMB)

**HISTAENS.rc.tmpl** - AGCM history for ensemble model integrations output

**mkiau.rc.tmpl** - controls generation of IAU-increment files used in ensemble of AGCM

**nmcperts.rc** - specifies information related to database of NMC-like perturbations

**mp_stats.rc** - defines configuration for calculation of ens-related statistics

**obs1gsi_mean.rc** - controls GSI observer mean

**obs1gsi_member.rc** - controls GSI observer members

**odsmatch.rc** - defines rules for conversion of GSI-diag files to ODS

**odsstats_ktonly.rc** - defines rules for summary of observation impact calculation

**post_egcm.rc** - defines which AGCM history output collections to calculate ens-statistics for

**satbias.acq** - specifies location of satellite (aircraft) bias coefficient files

The function of the `setup_atmens.pl` script briefly discussed in Sec. 3 is to properly set up these files and make sure the experiment environment is fully ready for use.

#### 4.1.4 Ensemble analysis options

The present GEOS Hybrid EVADAS implements three possibilities for generating an ensemble of analyses. The first follows Whitaker *et al.* (2008) and implements a square-root-type ensemble Kalman filter (EnSRF); this is the version currently used in the GMAO Forward Processing System (as well as in the settings of the global operational system at NCEP — though with differences in parameters and observer settings). The second procedure creates an ensemble of analyses by simply perturbing the central analysis with adequately scaled NMC-like perturbations. The third procedure creates an ensemble of GSI analyses, thus providing the environment to run an Ensemble of Data Assimilation (EDA) systems within GEOS (analogous to ECMWF's EDA; Isaksen *et al.* 2010). Of these options, only the first two have been extensively examined at GMAO thus far. The presence of certain resource files controls which of the options is active. Their specific configuration follows below.

- EnSRF (default). This option is triggered when the following resource are present in the directory determined by `$ATMENSETC`:

  1. `obs1gsi_mean.rc` - controls parameters for the observer mean.
  2. `obs1gsi_member.rc` - controls parameters for each observer member.
  3. `atmos_enkf.nml.tmpl` - controls options of atmospheric EnKF software itself.

  The presence of these resource files automatically triggers the sequence of events controlled by `obvsr_ensemble.csh`. The core of the observers' work is performed by GSI: the `obs1*.rc` files above correspond to selective configurations of the GSI namelist parameters which, among other things, tell GSI to bypass any minimization and invoke only the forward (nonlinear) obervation operators. In particular, the configuration in these resource files requests GSI to write out a set of diagnostic files containing corresponding OMB residuals and terms relevant to a subsequent EnKF(EnSRF) call. The member observers are told not to read the observations from scratch, but rather to read the observations that passed quality control and initial thinning as determined by the mean observer — this is why the member observers must wait for the mean observer to complete its task (see Fig. 24). The presence of `atmos_enkf.nml.tmpl` triggers the EnSRF, following the successful termination of the observers (again, see Fig. 24). The EnSRF normally uses background error localization parameters shared with the central (hybrid) GSI. It is possible to overwrite those by placing an alternative resource file[28] under `$ATMENSETC`. This might be useful depending on considerations related to resolution of the members versus that of the central GSI.

- Filter-free Ensemble Scheme. The scheme that creates an ensemble of analyses based simply on the central (hybrid) analysis and scaled NMC-like perturbations can be trigged by the presence of the following resource file:

  1. `easyana.rc` - trigger for the filter-free scheme; it contains specific information about the resolution of the central analysis and that of the members to be created.

  This procedure does not require running the observers, and it is meant to bypass the EnSRF. That is, in its most trimmed form, it allows for removal of the resource files associated with the observers and the EnSRF (as in the default settings seen above). However, leaving these files in place and simply adding the resource file `easyana.rc` triggers the Filter-free procedure, while still running the observers and the EnSRF, which at this point become simply diagnostic tools *not* feedbacking into the cycle. In order words, when `easyana.rc` is present in the `$ATMENSETC`, the Filter-free approach takes precedence. Using the Filter-free scheme while still running the observers and EnSRF is useful for testing and tuning.

- Ensemble of GSIs. This requires the following resource files to be placed in the `$ATMENSETC` directory:

  1. `gsi_mean.rc` - set to run only the mean observer (similar to obs1_mean.rc) – no minimization takes place. This is done so all member analyses (next) use the same quality-controlled set of observations.
  2. `gsi_member.rc` - controls how GSI analyzes each set of member backgrounds. The minimization options here can be set just as in the central analysis (without the hybrid option).

  To properly trigger the ensemble of GSIs — which turns the system into an Ensemble of Data Assimilation Systems — the resource files related to either the EnSRF or the Filter-free scheme must not be present in the directory `$ATMENSETC`. Users should be aware that though this provides GEOS ADAS with the capability to perform EDA, it is still a largely premature knob needing particular attention

[28] For a 1-degree ensemble this file would be `gmao_global_hybens_locinfo.x288y181l72.l72.rc`

when it comes to adequately perturbing observations, sea-surface-temperature, and model physics as in Isaksen *et al.* (2010).

### 4.1.5 Configuration of the ensemble of AGCMs

Once the user runs `setup_atmens.pl`, the ensemble of AGCMs is ready to cycle. The main resource files related to the model ensemble, found under `$ATMENSETC`, are `CAP.rc.tmpl`, `AGCM.rc.tmpl`, and `HISTAENS.rc.tmpl`, which control, respectively, the integration length and heartbeat of the model, general parameters in the model, and the list of output streams to be generated by the model. The comments that follow relate to resource files less typically known by users. These files are in general not a concern to most applications, but at times it might be helpful to know what they control:

**AGCM.BOOTSTRAP.rc.tmpl** – It is possible to bootstrap the AGCM initial condition (restart) files related to the ensemble of model integrations. This works in nearly the same way as the bootstrapping of restarts in the central ADAS. Just placing this file under the directory `$ATMENSETC` triggers bootstrapping. After the ensemble finishes one full cycle, a complete set of restarts and associated variables with be available for subsequent cycle, and the scripts have the function to rename this bootstrap resource file so it no longer affects further cycles.

**GEOS_ChemGridComp.rc** – This file is typically the same as that found under `$FVHOME/run`, meaning the ensemble of model integrations handle the same chemistry components as those used in the central model integrations. Early in the development of the hybrid system, tests were run with the ensemble model integrations not using aerosols from GOCART, thus requiring this file to differ from that used by the central ADAS model integrations. This is no longer the case, and in principle this file does not need to be present in `$ATMENSETC`.

**CAP_hh.rc.tmpl** – It might be desirable for different cycles to run the ensemble of model integrations out to different lengths of time. This can be done by placing a copy of the `CAP.rc.tmpl` file with a cycle starting time in it, as in `CAP_hh.rc.tmpl`, and having the `JOB_SGMT` parameter adjusted to control the length of model integration for cycles started at hour `hh`, differing from the default value of 12 hours. Recall that GMAO cycles start 3 hours off the synpotic times; thus valid times for `hh` are 03, 09, 15 and 21.

**HISTAENS_hh.rc.tmpl** – It is also possible to choose different model output streams from integrations starting at different times. This can be done by having files named `HISTAENS_hh.rc.tmpl` placed in `$ATMENSETC` for each of the non-conventional cycles; `hh` here follows the same convention as above.

### 4.1.6 Archiving output from ensemble ADAS

Archiving the output of the ensemble is controlled by the file `atmens_storage.arc` and by the ensemble ADAS environment variable `ENSARCH_FIELDS`. The former works just like any typical archiving resource file in GEOS ADAS and specifies template names for each type of file to be archived. For example, the file holding text output from the EnSRF is templated as `%s.atm_enkf.log.%y4%m2%d2_%h2z.txt`, and this is as it appears in `atmens_storage.arc`, preceded by the directory location where files are supposed to be placed in the archive, as for example, `${PESTOROOT}%s/atmens/Y%y4/M%m2`. The environment variable `${PESTOROOT}` is defined as in the regular ADAS configuration file.

The ensemble-specific environment variable `ENSARCH_FIELDS` defines collections of files generated in the ensemble ADAS to be packed together before archiving takes place. At the time of this writing, the following collections exist:

**eaaer** – set of aerosol analysis files; usually, file type `aana.eta`

**ebaer** – set of aerosol background files; usually, file type `abkg.eta`

**eana** – set of analysis files; usually, file type `ana.eta`

**easm** – set of assimilated files; usually, file type `asm.eta`

**ebkg** – set of background files; usually, file types `bkg.eta/sfc`

**ecbkg** – set of chemistry background files; usually, file type `cbkg.eta`

**edia** – set of model output, other than those in this list

**eoi0** – set of zero-hour observation impacts on mean ensemble analysis

**eniana** – set of non-inflated analysis produced by EnSRF

**eprg** – set of prognostic files; usually, file type prog.eta

**erst** – set of AGCM restarts; usually, the files with "bin" suffix

**evtk** – set of vortex tracks calculated for each ensemble member

**stat** – set of ensemble diagnostic files: mean, RMS, and spread

The collections are specified in `ENSARCH_FIELDS` separated by commas. A given collection determines a given type of file to be placed in a tar-file created before the usual `archive` script is launched. As mentioned earlier, the collections needed to reproduce or restart the ensemble at any time are "ebkg", "'cbkg", "erst", and "stat" (see Sec. 8.4). Note that any output generated by the ensemble of model integrations that does not fall in one of the recognized classes above ends up being saved as part of the "edia" class.

## 4.2 Deterministic (central) ADAS

This section presents specifics relevant to exercising the GMAO hybrid ADAS (the top grey-shaded box in Fig. 2). From a high level, there are only two main changes in the way GEOS ADAS works in hybrid mode as opposed to its more traditional (3D-Var-like; MERRA-2) mode. The first relates to settings of the GSI atmospheric analysis which must have its hybrid trigger turned on, choice of 3D or 4D minimization specified, corresponding linear balance option procedure selected, and other minor choices arranged. The second relates to how the GEOS forecasting model treats the analysis increment: when 3D-hybrid is selected in GSI, no special handling of the increments are required beyond what is done in traditional 3D-Var-like IAU mode; when 4D-hybrid is selected in GSI, the generation of 4D-increments from its minimization allows for the model to exercise various 4D flavors of IAU. These settings are more explicitly discussed in what follows.

### 4.2.1 Configuration of hybrid GSI

To trigger the hybrid option in GSI one must properly set parameters in the namelist `HYBRID_ENSEMBLE`, of the `gsi.rc.tmpl` file found in the `$FVHOME/run` directory. As with the setting up of the ensemble ADAS discussed in the previous section, typically there should be no need for a user to have to set these parameters on his/her own; when the setup of an experiment is complete (viz., `fvsetup`) all selections discussed here will be properly ready for an experiment to run. The purpose here is simply to highlight the essential parameters related to the hybrid system so users can experiment, develop and test at will with some adequate minimal knowledge.

A typical hybrid experiment, with the ensemble resolution at 1-degree, has the following GSI hybrid namelist settings:

```
&HYBRID_ENSEMBLE
  l_hyb_ens = .true.,n_ens = 32,generate_ens = .false.,
  uv_hyb_ens = .true.,s_ens_h = 800.,s_ens_v = -0.5,
  jcap_ens = 126,nlat_ens = 181,nlon_ens = 288,aniso_a_en = .false.,
  jcap_ens_test = 126,
  oz_univ_static = .true.,
  readin_localization = .true.,
  readin_beta = .true.,
  use_gfs_ens = .false.,
/
```

The parameter `l_hyb_ens` is the main trigger for the hybrid option in GSI. As indicated by the parameter `n_ens`, the default GMAO hybrid ADAS configuration uses a 32-member ensemble, with the resolution of its regular lat-lon grid specified by the parameters `nlat_ens` and `nlon_ens`.

An important feature of GSI hybrid relates to the possibility of providing different localization scales at different levels. This is controlled by the parameter `readin_localization`. Setting this parameter to *true* forces GSI to look for these scales in a file named `hybens_info`[29] in `$FVHOME/run` directory. Default localization scales have been tuned for two present resolutions of interest in GEOS. Another parameter to be aware of specifies the weights between the contribution from the ensemble error covariance and its climatological counterpart. The setting `beta1_inv = 0.5` appearing above suggests the contributions from these terms will be evenly divided. However, attention must be given to the parameter `readin_beta`. When this parameter is set to *true*, as in the example above, `beta1_inv` is ignored and GSI expects to read vertically-varying $\beta_c$ and $\beta_e$ parameters from the same file holding the localization scales, namely, "`hybens_info`". Figure 25 shows our current horizontal localization and "$\beta$" weights as a function of the vertical levels of the analysis. These settings give equal weights to each background error covariance term up to 5 hPa. Above this level, a transition layer is present where the weights slowly change so that above 1 hPa full weight is given to the climatological background error covariance matrix and no weight is given to the ensemble contribution (similar in nature to Clayton *et al.* (2013); see Fig. 7 in that work). The other parameters appearing in the resource file (namelist) above can basically be ignored by most users.

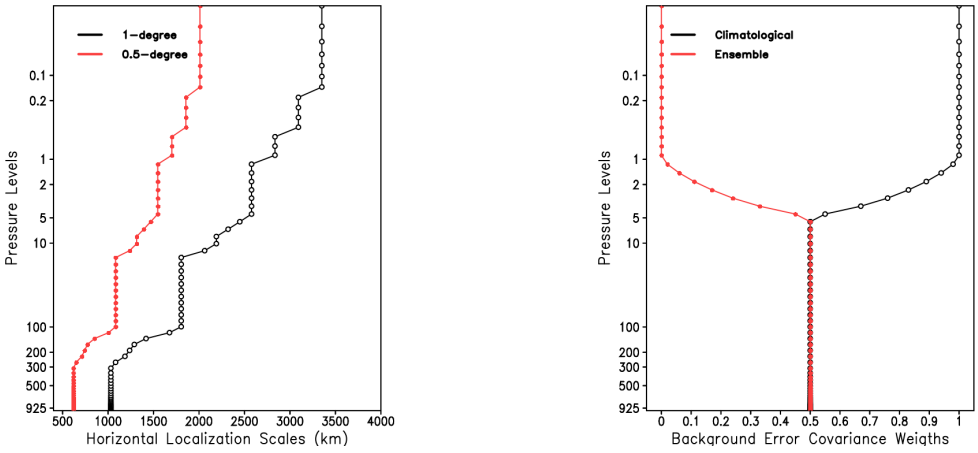


**Figure 25:** Left panel: horizontal localization scales (km) for two model resolutions: $1^o$ ($\sim$ 100 km; black) and $0.5^o$ ($\sim$ 50 km; red). Right panel for (dimensionless) vertical weights given to climatological (black) and ensemble (red) background error covariances. Values in both panels shown as a function of analysis pressure levels.

There are a number of namelists inside `gsi.rc.tmpl`, each responsible for controlling specific features in GSI. Users are referred to the GSI online documentation (Hu *et al.* 2016) for more details. However,

[29]For example, the file `gmao_global_hybens_info.x288y181l72.rc` for a 1-degree ensemble

two sets of parameters are of particular interest here as they refer specifically to knobs related to the hybrid choice in the analysis system. First, the namelist `STRONGOPTS` controls choices related to the GSI TLNMC balance constraint (see Kleist *et al.* 2009b). Specifically, the parameter `hybens_inmc_option` controls how the balance constraint is applied to the increments generated by GSI; the choice depends on whether GSI is set as a 3D or 4D solver. In a Hybrid 3D-Var this parameter is set to 2; in a Hybrid 4D-EnVar setting, the best choice is 3, which applies the TLNMC to each of the 4D-incremental corrections estimated by the GSI minimization. Both these options will balance the climatological and ensemble contributions to the increment [both terms on the right-hand-side of (3)].

The second set of parameters of significance to this document are the options in GSI controlling its hybrid 3D and 4D configurations. The `SETUP` namelist, in `gsi.rc.tmpl`, contains the main option controlling this setting. The logical parameter `l4denvar` triggers the (Hybrid) 4D-EnVar, when set to *true*. When this parameter is left out (i.e., set to *false*), GSI runs Hybrid 3D-Var as long as the parameter `l_hyb_ens`, seen earlier, is set to true; otherwise, traditional 3D-Var is set.

### 4.2.2 Configuration of atmospheric GCM IAU forcing

As mentioned in the introduction, when GSI is set to run in 4D mode, it produces incremental corrections to the background at desired frequency. Availability of high-frequency 4D increments allows for more frequent IAU updates following the strategies of 4DIAU. Various flavors of 4DIAU are controlled by the AGCM; see Takacs *et al.* 2018 for details. In any hybrid experiment within GEOS ADAS the user will find two copies of the file `AGCM.rc.tmpl`: one residing under `$FVHOME/run`, responsible for controlling the deterministic (hybrid) model integration; and another residing under `$ATMENSETC`, responsible for controlling the model integration of the ensemble members. The one related to the deterministic DAS will have parameter settings related to 4DIAU as follows:

```
REPLAY_MODE: Exact_4D
REPLAY_FILE: agcm_import_rst.%y4%m2%d2_%h2z.bin
CORRECTOR_DURATION:  3600
REPLAY_SHUTOFF: 21600.0
```

The `REPLAY_MODE` variable distinguishes between 4DIAU and (3D)IAU; `REPLAY_FILE` defines the templated name of the file contaning the IAU increment (not a tendency at this point); `CORRECTOR_DURATION` defines the length of the IAU-corrector step, which in the 4D case is typically set to 1 hour (in seconds); and finally, the parameter `REPLAY_SHUTOFF` tells the model when to stop looking for files containing analysis increments. This latter is associated with the length of the assimilation window, which is presently set to 6 hours (in seconds). Since the assimilation strategy for the ensemble members is currently 3D-EnSRF, the corresponding model integrations run the 3D (traditional) version IAU. Whichever flavor of IAU the model is set to run includes the internal default of modulating the analysis tendencies with a digital filter — in 3D this essentially picks up from the ideas in Polavarapu *et al.* (2004); Takacs *et al.* (2018) lay out the details for how to apply the digital filter in 4D.

This page intentionally left blank.

# 5 Scheduler

As mentioned earlier, the present way of running the hybrid system begins with the user submitting the usual `g5das.j` script to the batch system and letting it submit the ensemble ADAS job script, `atm_ens.j`. Under the *scheduler*, the user submits instead a job script called `EXPID_scheduler.j`, where `EXPID` stands for the user experiment name. This job is responsible for controlling how the central and the ensemble ADAS executions are coordinated. This mode of running is aimed at maximizing efficiency and exploiting most of the parallelism existing between the central ADAS and the ensemble ADAS (see Fig. 2). Recall that these two systems are coupled to each other in two ways: (i) the central ADAS GSI analysis requires the ensemble of background states; and (ii) the ensemble ADAS requires the central GSI analysis for the re-centering of its analysis members [30]. This means that as soon as the central GSI analysis is finished, the ensemble ADAS can begin its task. In other words, the ensemble ADAS can run concurrently with the 12-hour (or longer) initialized forecast of the central ADAS. Once these two components are running in parallel, the scheduler is responsible for synchronizing the two systems. A new central analysis can only start once the ensemble ADAS has produced the necessary ensemble of background fields. At the time of this writing only a preliminary version of the scheduler exists; therefore we leave further description to a later revision of this manuscript.

In this mode of running, when the GSI hybrid and the ensemble (EnKF) analyses are concurrent, the satellite biases needed by the ensemble analysis must come from a previous cycle, thus being the same as used by GSI. In this case, the `satbias.acq` file under `FVHOME/run/atmens` must point to the directory defined by the environment variable `RSTSTAGE4AENS`. An example of such `satbias.acq` is

```
/discover/nobackup/USER/EXPID/atmens/RST/EXPID.ana_satbias_rst.%y4%m2%d2_%h2z.txt  = >
           EXPID.ana.satbias.%y4%m2%d2_%h2z.txt
/discover/nobackup/USER/EXPID/atmens/RST/EXPID.ana_satbang_rst.%y4%m2%d2_%h2z.txt  = >
           EXPID.ana.satbang.%y4%m2%d2_%h2z.txt
```

for the case when `RSTSTAGE4AENS` is set to `/discover/nobackup/USER/EXPID/atmens/RST`.

If the job needs to be re-started from the beginning, say for the cycle on 1776070415, simply remove the work directories `fvwork_EXPID_1776070415` and `enswork_EXPID_1776070415`, and resubmit the scheduler with the refresh option, that is,

```
qsub -v refreshdate = 1776070415 EXPID_scheduler.j
```

Following the framework developed to control completion status of various procedures for the EnADAS, the scheduler has its own set of hidden variables. The following is a list of relevant variables used in `edas_scheduler.csh`:

**.DONE_MEM001_rstcp.1776070415** :
signals successful copy of relevant restart files from `recycle` directory to directory defined by `RSTSTAGE4AENS`

**.DONE_MEM001_analyzer.1776070418** :
signals successful completion of central (hybrid) analysis

**.DONE_MEM001_atm_ens_eana.1776070415** :
signals successful completion of ensemble analysis (presently, EnKF)

**.DONE_MEM001_atm_ens.1776070415** :
signals successful completion of EnADAS

**.DONE_MEM001_ddas.1776070415** :
signals successful completion of central ADAS

[30] Currently, another link exists since we use the satellite bias estimates of the central analysis to start the same-synoptic-time ensemble analysis. This can be relaxed by having the ensemble of observers use the previous-cycle satellite bias estimates instead.

By construction, all the hidden files appear under the `FVHOME` directory.

An *important* note refers to stopping a cycling experiment. One easy way to momentarily interrupt a cycling experiment is to simply move the file holding the batch script to another, temporary name. For example, in a purely deterministic ADAS case, the user can prevent the cycling from continuing by simply renaming the script `g5das.j` to something else, e.g., `g5das.j.hold`. This way, when the ongoing job tries to re-submit itself, it will not find the batch job and will come to a halt. Similarly, in hybrid mode, the cycle can be interrupted by renaming the driving script, but in this case we must remember there are two scripts at play. The central ADAS, does not re-submit itself (`g5das.j`) but instead submits the ensemble ADAS script `atm_ens.j`, and cycling amounts to `atm_ens.j` submitting `g5das.j`. Therefore it is the script lined up to be next that must be renamed. If the central ADAS is the one running, then `atm_ens.j` must be renamed; if the ensemble is what is running, than `g5das.j` must be renamed. This is all fine without the use of the scheduler. However, when the scheduler is used to control the batch job submission, the user must be aware to *never rename the scheduler script*, `EXPID_scheduler.j`, for the purposes of interrupting the cycle. In this case, it is still the script `g5das.j` and `atm_ens.j` that must be renamed. Still, even here, the *user must exercise caution*. Since `atm_ens.j` is called twice by the scheduler, this job script can only be renamed after the second call has been made, i.e., after the ensemble of AGCM forecasts has been launched.

# 6 General Sanity Checks Recommended to Users

Just as with GSI, the log file of the EnKF echoes out a table of observation-minus-background (OMB) and observation-minus-analysis (OMA) residuals. One expects that a reasonably well configured system will have the GSI tables and those from the EnKF looking rather comparable, particularly if the thinning data strategy between the central analysis and that of the ensemble analysis is kept the same.

The following is an example of the OMB ($J_o$-observation fit) table from the central hybrid GSI analysis:

| Observation Type | Nobs | Jo | Jo/n |
|---|---|---|---|
| surface pressure | 58993 | 7.7543508538588003E+03 | 0.131 |
| temperature | 100202 | 1.4151903304701959E+05 | 1.412 |
| wind | 425416 | 4.1748592693636852E+05 | 0.981 |
| moisture | 14011 | 8.4253060120348000E+03 | 0.601 |
| ozone | 17149 | 3.1765657511636487E+04 | 1.852 |
| gps | 54419 | 8.7688179856585470E+04 | 1.611 |
| radiance | 2487638 | 4.0790375455413770E+05 | 0.164 |
| | Nobs | Jo | Jo/n |
| Jo Global | 3157828 | 1.1025422087716414E+06 | 0.349 |

and below is an example of the table from the EnKF:

| Observation Type | Nobs | Jo | Jo/n |
|---|---|---|---|
| surface pressure | 58966 | 8.9479736328125000E+03 | 0.152 |
| temperature | 100154 | 1.4063432812500000E+05 | 1.404 |
| wind | 425398 | 4.2579856250000000E+05 | 1.001 |
| moisture | 14004 | 7.3794936523437500E+03 | 0.527 |
| ozone | 18875 | 3.1167117187500000E+04 | 1.651 |
| gps | 54451 | 8.5983710937500000E+04 | 1.579 |
| radiance | 2492015 | 4.3177943750000000E+05 | 0.173 |
| Jo Global | 3163863 | 1.1316906250000000E+06 | 0.358 |

Both of these are for the analysis at 0000 UTC, on 1 April 2012. The hybrid analysis uses an ensemble that has already been span up. We see, for example, that the EnKF is not taking precipitation observations. This is done by construction; we choose not to take this data-type for now since the EnKF requires a better handle on this type of information. For most data-types, the difference in observation count ranges from a few dozen to a few hundred at most. This is largely attributed to quality control decisions and the difference between using instantaneous 0.5-degree resolution backgrounds in the hybrid GSI analysis and 1-degree ensemble mean backgrounds in the EnKF analysis case. The most noticeable difference in the two tables above comes from comparing the radiance observation counts. We see that the EnKF ends up taking 4377 radiance observations more than the hybrid analysis. This is attributed to the fact that in our implementation the satellite bias correction coefficients used by the EnKF come from the central hybrid analysis. That is, the EnKF only executes after the central analysis has finished its work. Though we can use the satellite bias estimates from the previous cycle, as the central hybrid GSI analysis does, we take advantage of the availability of the current estimates from the central ADAS when running the EnKF analysis. Most importantly, when comparing the two tables above, we see the similarity between the observation fits scaled by the number of observations ($J_o/n$, last column). It is difficult to establish a rule of thumb for how these numbers should compare, other than wanting them to be close. That the EnKF scaled fits are so close to those from the central analysis is interesting given that the former observer works from an ensemble mean state which is not necessarily physical. Recall from the discussion in Sec. 1.2.1 that the count of the observations taken in by the EnKF/EnSRF is not indicative of the count of observations *actually* analyzed.

Similarly, the observation fits to the analysis from the central GSI look as below:

| Observation Type | Nobs | Jo | Jo/n |
|---|---|---|---|
| surface pressure | 59046 | 4.9852106626854447E+03 | 0.084 |
| temperature | 100203 | 7.7757153869908128E+04 | 0.776 |
| wind | 429436 | 2.5802116043830608E+05 | 0.601 |
| moisture | 14011 | 4.9607887271425416E+03 | 0.354 |
| ozone | 17149 | 8.9490519057384627E+03 | 0.522 |
| gps | 55560 | 5.0622628146044233E+04 | 0.911 |
| radiance | 2569945 | 3.3927362893704593E+05 | 0.132 |
| | Nobs | Jo | Jo/n |
| Jo Global | 3245350 | 7.4456962268687086E+05 | 0.229 |

with the equivalent (a posterior fits) table from the EnKF looking as:

| Observation Type | Nobs | Jo | Jo/n |
|---|---|---|---|
| surface pressure | 58966 | 6.0409160156250000E+03 | 0.102 |
| temperature | 100154 | 9.5538234375000000E+04 | 0.954 |
| wind | 425398 | 3.1510571875000000E+05 | 0.741 |
| moisture | 14004 | 5.6834106445312500E+03 | 0.406 |
| ozone | 18875 | 1.9777376953125000E+04 | 1.048 |
| gps | 54451 | 5.1486160156250000E+04 | 0.946 |
| radiance | 2492015 | 3.8114400000000000E+05 | 0.153 |
| Jo Global | 3163863 | 8.7477581250000000E+05 | 0.276 |

Again, the two tables are very comparable. Indeed the reduction in $J_o/n$ for both analyses is very similar, though the central hybrid analysis tends to fit the observations slightly more closely.

Another illustration, for the same synoptic time, is given below, where now the global fits to the background are illustrated for the central analysis (same as total shown in the first table above), the mean observer, and the components of a 32-member ensemble of observers. Comparing the count from the EnKF fits to the prior, we see that the EnKF takes in only slightly fewer observations than does the mean observer – this is due to a consistency check within the EnKF software.

| | | | | |
|---|---|---|---|---|
| Central analysis | Jo Global | 3157828 | 1.1025422087716414E+06 | 0.349 |
| | | | | |
| obs_ensmean | Jo Global | 3167079 | 1.1400739327238461E+06 | 0.360 |
| obs_mem001 | Jo Global | 3123256 | 1.4146527504929921E+06 | 0.453 |
| obs_mem002 | Jo Global | 3128285 | 1.3512821520852332E+06 | 0.432 |
| obs_mem003 | Jo Global | 3131361 | 1.3420369419540870E+06 | 0.429 |
| obs_mem004 | Jo Global | 3125948 | 1.3649793228603806E+06 | 0.437 |
| obs_mem005 | Jo Global | 3121210 | 1.3938713409115009E+06 | 0.447 |
| obs_mem006 | Jo Global | 3131823 | 1.3535254403115206E+06 | 0.432 |
| obs_mem007 | Jo Global | 3125040 | 1.4356407879127199E+06 | 0.459 |
| obs_mem008 | Jo Global | 3122812 | 1.3893385599306291E+06 | 0.445 |
| obs_mem009 | Jo Global | 3132624 | 1.3556867770441249E+06 | 0.433 |
| obs_mem010 | Jo Global | 3122710 | 1.3602509368970357E+06 | 0.436 |
| obs_mem011 | Jo Global | 3126436 | 1.3569395372827167E+06 | 0.434 |
| obs_mem012 | Jo Global | 3126090 | 1.3789860104918096E+06 | 0.441 |
| obs_mem013 | Jo Global | 3132534 | 1.3629284646751210E+06 | 0.435 |
| obs_mem014 | Jo Global | 3125473 | 1.4240120201202775E+06 | 0.456 |
| obs_mem015 | Jo Global | 3124538 | 1.3686023277420180E+06 | 0.438 |
| obs_mem016 | Jo Global | 3125313 | 1.3625097896574179E+06 | 0.436 |
| obs_mem017 | Jo Global | 3121705 | 1.4513418843554165E+06 | 0.465 |
| obs_mem018 | Jo Global | 3124688 | 1.3476942535456968E+06 | 0.431 |

```
obs_mem019       Jo Global   3129546    1.3462618226212750E+06   0.430
obs_mem020       Jo Global   3132273    1.3767084034834011E+06   0.440
obs_mem021       Jo Global   3132648    1.3697893025347698E+06   0.437
obs_mem022       Jo Global   3126265    1.3684473016587161E+06   0.438
obs_mem023       Jo Global   3126355    1.3960081508113015E+06   0.447
obs_mem024       Jo Global   3127770    1.3681285698646517E+06   0.437
obs_mem025       Jo Global   3122046    1.4015666052547472E+06   0.449
obs_mem026       Jo Global   3120662    1.4080196645814902E+06   0.451
obs_mem027       Jo Global   3124172    1.3628864136518587E+06   0.436
obs_mem028       Jo Global   3126496    1.3329366092685640E+06   0.426
obs_mem029       Jo Global   3132788    1.3599484934718781E+06   0.434
obs_mem030       Jo Global   3113985    1.4292567105894105E+06   0.459
obs_mem031       Jo Global   3117175    1.3533330934446647E+06   0.434
obs_mem032       Jo Global   3123969    1.3742827905094966E+06   0.440
```

Though the observations taken in by the individual observers are the same as those taken in by the mean observer, there is still a level of check done by each observer that discards observations not sufficiently close to its respective set of backgrounds. As it turns out, individually, a given observer allows for slightly fewer observations to be used than the mean; another way to say this is that each member background provides a slightly worse fit to the observations than the ensemble mean does (compare also the $J_o/n$ columns). In fact, a theoretical argument for the linear case (when the observation set is kept constant) can be made to better explain this finding. If we denote $\bar{J}_o$ to be the observer mean fit, it is easy to show that

$$< J_{o;m} >= \bar{J}_o + \frac{1}{M}\sum_{m=1}^{M}\sum_{i=1}^{p}(\mathbf{x}_m - < \mathbf{x}_m >)^T \mathbf{H}_i^T \mathbf{R}_i^{-1} \mathbf{H}_i (\mathbf{x}_m - < \mathbf{x}_m >)\,, \tag{13}$$

where $p$ is the total number of observations, $M$ is the total number of ensemble members, $< \bullet >$ denotes the ensemble average operator, and $J_{o;m}$ stands for the observation fit (cost) evaluated for the $m$-th member of the ensemble. Noticing the second term on the right is positive definite, we should always expect the observation fit to the mean ensemble backgrounds to be smaller than the average of the ensemble of observer fits to the observations. This is precisely what we see in the comparison above. Neglecting the difference in the observation count, we see the observation fit to the mean to be roughly $1.14E+06$ and the average of the observers observation fits to be roughly $1.37E+06$. It is useful to notice that application of the trace operation to (13) leads to the alternative expression

$$\begin{aligned} < J_{o;m} > \quad &= \quad \bar{J}_o + \frac{M-1}{M} Tr\left(\mathbf{B}_e \sum_{i=1}^{p} \mathbf{H}_i^T \mathbf{R}_i^{-1} \mathbf{H}_i\right) \\ &\overset{M\to\infty}{\longrightarrow} \quad \bar{J}_o + Tr\left(\mathbf{B}_e \sum_{i=1}^{p} \mathbf{H}_i^T \mathbf{R}_i^{-1} \mathbf{H}_i\right) \end{aligned} \tag{14}$$

which states that the averaged observers fit is always biased with respect to the observer fit to the mean, regardless of the size of the ensemble. This remains so even when the ensemble background error covariance matrix $\mathbf{B}_e$ becomes an accurate representation of the true background error covariance matrix.

Another way to corroborate desirable behavior from the hybrid system is to examine the observation counts and fits as the assimilation progresses. A comparison of the hybrid analysis with a control (traditional 3D-Var) experiment at initial time (0000 UTC on 1 April 2012) shows:

```
Standard Analysis
           Jo Global        3157828    1.1025422087716414E+06       0.349
           Jo Global        3220086    7.6636349426764739E+05       0.238
           Jo Global        3242795    7.5913357050380018E+05       0.234
Hybrid Analysis:
           Jo Global        3157828    1.1025422087716414E+06       0.349
```

```
           Jo Global        3223267    7.5266204694107187E+05       0.234
           Jo Global        3245350    7.4456962268687086E+05       0.229
```

where we see that initially both the standard and hybrid analyses begin from the same background (notice the same initial data count and observation fits). As the minimization progresses, the standard analysis ends up taking in fewer observations than the hybrid analysis; this is an ideal behavior. At this initial stage it would be acceptable, though not ideal, for the hybrid analysis to take in slightly less data. However, as time progresses this would not be acceptable. Indeed, in the experiments used for the present illustration, a few cycles away from the initial cycle, at 0000 UTC on 2 April 2012, the situation is rather clear, as shown below:

```
 Standard Analysis
           Jo Global        3101111    1.1045415299101665E+06       0.356
           Jo Global        3161479    7.6260731657937751E+05       0.241
           Jo Global        3184423    7.5556159453427303E+05       0.237
 Hybrid Analysis:
           Jo Global        3109514    1.0803129941124925E+06       0.347
           Jo Global        3170399    7.4634125953013694E+05       0.235
           Jo Global        3190465    7.3919985174608813E+05       0.232
```

Before the start of the minimization, the hybrid analysis already comes in with more observations than the control (standard) cycle. More impressive is the fact that even with the increased number of observations, the global observation fits (second column of numbers) shows smaller numbers than does the control analysis. Ideally we want more observations, smaller fits, and smaller $J_o/n$; this is exactly what we see happening with the hybrid analysis in the case illustrated above.

# 7 Auxiliary Programs

## 7.1 Additive inflating perturbations

Among the set of peripheral components necessary to maintain the ensemble of analyses is the procedure allowing for generation of random perturbations used for additive inflation. Presently, there are two ways in which GEOS can generate these fields:

**GSI internally-generated perturbations.** It is possible to re-configure the GSI-observer mean to write out as many randomly generated perturbations as members of the ensemble. This can be done by modifying the resource file `obs1gsi_mean.rc`. Perturbations generated in this way are normally distributed with zero-mean and with error covariance structures derived from the GSI climatological background error covariance matrix, that is, $\mathscr{N}(\mathbf{0}, \mathbf{B}_s)$. We have experimented with using these perturbations to inflate the ensemble and have found the ensemble to collapse rather quickly in this case. In other words, it seems that these random perturbations are simply noise that the model integrations quickly dissipate.

**NMC-like perturbations.** Alternatively, following the idea of J. S. Whitaker and what is implemented operationally at NCEP, we generate perturbations for additive inflation by selecting randomly from a year-long database of NMC-like perturbations. At GMAO, the database has been constructed from GEOS 5.7 year-long forecasts. These perturbations are identical to those used to tune the GSI climatological background error covariance matrix presently used in our traditional 3D-Var, that is, the perturbations consist of differences from the relevant 48- and 24-hour forecast fields (zonal and meridional winds, virtual temperature, specific humidity, surface pressure, and ozone). The location of the database is found (and specified) in the resource file `nmcperts.rc`.

The present default setting of the ensemble ADAS randomly selects from the database perturbations as many files as members. The selection is season-aware, in the sense that perturbations are chosen for the season associated with the experiment's current analysis time. At any given analysis time, the ensemble ADAS scripts will calculate the mean of the selected perturbations and create a new set of perturbations with mean removed. Recall that perturbations in the database are for a specific year and are likely not related to the time of the analysis cycle of any one experiment. In some instances, depending on the procedure used to calculate and remove the perturbations mean, the new set of perturbations might be written out to file at the resolution of the ensemble, rather than at their original resolution of 0.25 degrees. The presence of the resource file `mp_stats_pert.rc` in the directory defined by `ATMENSETC` allows for the perturbations to be converted to low resolution (see Sec 7.3).

## 7.2 Ensemble re-centering and inflation

Another fundamental component of the ensemble ADAS is the part handling re-centering. This initial implementation of the ensemble ADAS only handles the usual meteorological fields: winds, temperature, specific humidity, surface pressure, and ozone. These fields are present in the background and analysis files of each ensemble member. Those familiar with GEOS ADAS will know that the files carrying these fields form the basis of the so-called *dyn-vector*. The re-centering program is based on the module `m_dyn.f90` of the `GMAO_hermes` library. The is convenient since it allows for use of the various dyn-capabilities, such as automatic interpolation and remapping due to topography changes. As with all other dyn-based programs, the so-called `dyn_recenter.x` is command-line driven, and its specific usage is shown in Fig. 26.

In its most basic form, the program requires an input file containing fields from a given member, another containing the ensemble mean, and a third file containing the analysis to re-center about, typically coming from the hybrid (central) ADAS. Without any other input, the re-centering program will overwrite the input ensemble member file. In the dual-resolution configuration of the hybrid, any given ensemble member and its corresponding ensemble mean are at lower resolution than the hybrid analysis to center about. The re-centering program automatically interpolates the central analysis to the resolution of the member. This must

```
     ------------------------------------
     dyn_recenter - recenter ensemble mean
     ------------------------------------


Usage:

  dyn_recenter.x [options]
                 x_e(i) x_m x_a [-o output file]

where [options]

-h              Help (optional)
-g5             Treats files as GEOS-5 files
-damp           Apply damp to levels above 5mb
-noremap        Force no-remap whatsoever
-remap2central  Remap member and ensmean to central
                (Default: remap central and ensmean to member
-a       factor Multiplicative factor for inflating perturbations
-inflate fname  filename containing inflating perturbations
-verbose        echoes general information
-o       fname  filename of resulting recentered fields
                (CAUTION, default: overwrite x_e(i) file)

 Required inputs:
  x_e(i)  - filename of field to be recenter
  x_m     - original mean
  x_a     - new mean around which member gets recentered

 Remarks:
  1. This program is used in context of ensemble DAS to recenter
     dyn-vector around desired mean. That is, assuming the ensemble
     mean is x_m, and a desired center mean is x_a, this program
     reads multiple members x_e of an ensemble of dyn-vectors
     and calculates: x_e(i)  =  x_e(i)) - x_m + x_a, for ensemble
     member i.
  2. There are a million ways to write a more efficient code
     for this - indeed one might need to do this using ESMF
     to better handle high-resolution fields.
```

**Figure 26:** Command line for re-centering program used not only to re-center ensemble analyses about central hybrid analysis, but also to apply additive inflation, vertical blending, and remapping.

be done with care for the change in topography. By default, the program remaps the central high-resolution analysis to the topography of the member analysis. Flags exist to either turn remapping off or remap in the other direction.

This program is also responsible for handling the additive inflation procedure. With properly specified inputs, the program `dyn_recenter.x` will also add a scaled perturbation field, properly interpolated (if necessary), to the member. Furthermore, at present, the ensemble ADAS runs with vertical blending applied to its members, so that between 20 and 5 hPa, any member analysis is smoothly merged into the central analysis. That is, above 5 hPa, the ensemble has no variance. This is consistent with what we discussed in Sec. 4.2.1 and with our present choice of vertically-varying weights given to both the climatological and ensemble background error covariances in the hybrid GSI, namely, with GSI relying solely on its climatological background errors in the high atmosphere.

## 7.3 Ensemble mean and RMS

Use of an EnKF-based ensemble assimilation strategy requires, at the very least, calculation of the ensemble mean. This is needed to form quantities such as $\mathbf{h}(\mathbf{x}_m) - \mathbf{h}(\bar{\mathbf{x}})$. In practice, these differences come from the difference between the OMB formed by the individual observer members and the OMB formed by the mean observer. This is why the ensemble ADAS schematic shown in Fig. 24 displays a box corresponding to the *observer mean*. The other place where the ensemble mean is required is during the re-centering step, when it is actually the mean of the ensemble of analyses that is required (see previous subsection).

In GEOS ensemble ADAS there are a few ways to calculate the required ensemble mean. One uses the program `GFIO_mean.x`, found in the `bin` directory of a GEOS build. Another relies on the the program `mp_stats.x`, which is an ESMF-MAPL-based program; and yet a third (and by far the fastest) program is `mp_dyn_stats.x`, which uses the same strategy for reading the ensemble members as that used by the EnKF software. The latter two programs are also more efficient in that they are capable of producing the diagnostic produced by `pertenergy.x` and thus allow bypass of this secondary call by the scripts. Any of these programs is responsible for generating the ensemble mean, spread and the total-energy-scaled spread.

It might be convenient to know that programs like `mp_stats.x` are also capable of performing a number of other tasks. A closer look at its usage line is provided in Fig. 27.

Indeed, in Sec. 4.1.3, we have come across the resource files `mp_stats.rc` and `mp_stats_perts.rc` associated with running this program in different circumstances. The contents of these resource files control MPI distribution and resolution options. The two files just mentioned differ essentially in the way resolution is treated. The first one, related to calculation of required ensemble statistics, has resolution of its inputs and outputs equally set to the resolution of the ensemble members; the second one, used to de-bias the NMC-like perturbations, has its input set to the $0.25^o$ resolution of the perturbations and its output set to the $1^o$ (default) resolution of the members.

## 7.4 Energy-based ensemble spread

As just mentioned above, available automatic diagnostics being produced from the ensemble of analyses and backgrounds correspond to measures of the ensemble spread. These provide guidance for the reliability of the ensemble. Both root-mean-square error (with respect to the mean of the ensemble) and an energy-based RMS error measure are available as diagnostics. As just seen in the previous subsection, the default program used to calculate these diagnostics is `mp_stats.x`. The following defines the energy-based ensemble spread as calculated within this program (similar to `pertenergy.x`):

$$e = \sum_{m=1}^{M} \mathbf{e_m}^T \mathbf{E} \mathbf{e}_m \tag{15}$$

where the error vectors $\mathbf{e}_m = \mathbf{x}_m - \bar{\mathbf{x}}$, for each member $m$, and the matrix $\mathbf{E}$ is taken as a linearized form of the total energy operator. That is, an energy-based deviation of each ensemble member from the mean can

```
Usage: mp_stat.x [options] files

options:

-o     FILE       specify output filename
-alpha NUMBER     specify multiplicative coeff to scale mean
                    before adding result to each file read in
                    (see -tmpl)
-date  NYMD NHMS  date/time of output file(s)
                    (when absent use date of last file read)
-ene   FILE       spefify output file containing energy-based
                    measure (NOTE: this triggers the calculation)
-etmpl ENEFTMPL   template for individual energy estimates for each member
                    (e.g., -etmpl myenergy.%y4%m2%d2_%h2z)
-nonrecene        de-activate recursive calculation of energy measure
                     (NOTE: this will sweep through the data twice)
-rms              calculate rms
-stdv  FILE       provide filename of output stdv
-umean FILE       provide available estimate of mean
                    (only used in non-recursive calcuation
                     of standard deviations)
-tmpl  FNAMETMPL  specify filename template of output files
                    NOTE: do not provide filename extension
                          (nc4 will be appended to name)
                    (e.g., -tmpl myfiles.%y4%m2%d2_%h2z)
-vars  LIST       where LIST is a list of variable separate


Example usage:

 1. Obtain mean:
    mp_stats.x -o mean.nc4 mem0*/hy05a.bkg.eta.20120410_00z.nc4

    1a. calculating monthly means (i.e., files from diff times)
        can be done by specifying date of output file, e.g.,
    mp_stats.x -o apr_mean.nc4 -date 20120401 0 mem0*/hy05a.bkg.eta.201204*z.nc4

 2. Recursively obtain rms subtracting user-specified mean from original fields:
    mp_stats.x -o rms.nc4 -rms mem0*/hy05a.bkg.eta.20120410_00z.nc4

 3. Recursively obtain stdv subtracting user-specified mean from original fields:
    mp_stats.x -o mean.nc4 -stdv stdv.nc4 mem0*/hy05a.bkg.eta.20120410_00z.nc4

 4. non-recursive stdv calc can be triggered by:
    mp_stats.x -o stdv.nc4 -usrmean mean.nc4 -rms -stdv NONE mem0*/hy05a.bkg.eta.20120410_00z.nc4
    in this case, the file mean.nc4 is an input such as that obtained w/ (1)

 5. non-recursive calc of energy-based error:
    mp_stats.x -usrmean mean.nc4 -ene ene.nc4 mem0*/hy05a.bkg.eta.20120410_00z.nc4
    in this case, the file mean.nc4 is an input such as that obtained w/ (1)

 6. removing mean from samples and writing out anomalies:
    mp_stats.x -tmpl anomaly.%y4%m2%d2_%h2z -alpha -1.0 -date 19990101 0 hy05a.ana.eta.*

 7. calculate energy-measure wrt to mean and write out individual member energy-measure estimates:
    mp_stats.x -nonrecene -ene mean_ene.nc4 -etmpl energy.%y4%m2%d2_%h2z mem0*/hy05a.bkg.eta.20120410_00z.nc4
       where: mean_ene.nc4   is ouput containing mean energy
           templated-files are ouput files containing each member's energy

 8. calculate energy-measure wrt central analysis and write out individual member energy-measure estimates:
    mp_stats.x -nonrecene -etmpl energy.%y4%m2%d2_%h2z -usrmean central_ana.nc4
               -date 20120410 0 mem0*/hy05a.bkg.eta.20120410_00z.nc4
       where: central_ana.nc4   is input central field
           templated-files     are ouput files containing each member's energy wrt to central

 9. removing mean of energy fields:
    mp_stats.x -alpha -1.0 -tmpl energy_anomaly.%y4%m2%d2_%h2z -usrmean mean_energy.nc4 energy.20120410_00z.*.nc4
```

**Figure 27:** Command line for `mp_stats.x` program. In its most basic use, this program performs similar calculations as those done by `GFIO_mean.x` and `mp_dyn_stats.x`, but with expanded diagnostic capabilities.

be evaluated using either of the following expressions (see Lewis et al. 2001; Errico et al. 2007):

$$e_t \equiv \mathbf{e}_m^T \mathbf{E}_t \mathbf{e}_m = \tfrac{1}{2} \sum_{i,j,k} \Delta H_{i,j} \Delta\sigma_{i,j,k} \left[ u_1' u_2' + v_1' v_2' + \frac{c_p}{T_r} T_1' T_2' + \frac{RT_r}{p_r^2} p_{s1}' p_{s2}' \right]_{i,j,k}, \tag{16}$$

$$e_v \equiv \mathbf{e}_m^T \mathbf{E}_v \mathbf{e}_m = \tfrac{1}{2} \sum_{i,j,k} \Delta H_{i,j} \Delta z_{i,j,k} \left[ u_1' u_2' + v_1' v_2' + \frac{c_p}{T_r} T_1' T_2' + \frac{RT_r}{p_r^2} p_{s1}' p_{s2}' \right]_{i,j,k}, \tag{17}$$

where $\Delta H_{i,j}$ is a horizontal grid-box weight and the distinction between the two norms is in how they weigh the fields in the vertical, with $\Delta\sigma_{i,j,k}$ and $\Delta z_{i,jk}$ being fractional weights, respectively, defined as:

$$\Delta\sigma_{i,j,k} = \frac{\Delta p_{i,j,k}}{p_{s,i,j} - p_t}, \tag{18}$$

$$\Delta z_{i,j,k} = \frac{\Delta \ln p_{i,j,k}}{\ln p_{s,i,j} - \ln p_t}. \tag{19}$$

The physical scaling coefficients $c_p = 1004.6$ J kg$^{-1}$ K$^{-1}$, $R = 287.04$ J kg$^{-1}$ K $^{-1}$, $T_r = 280$ K, and $p_r = 1000$ hPa are, respectively, the specific heat at constant pressure, the gas constant of dry air, and a reference temperature and pressure.

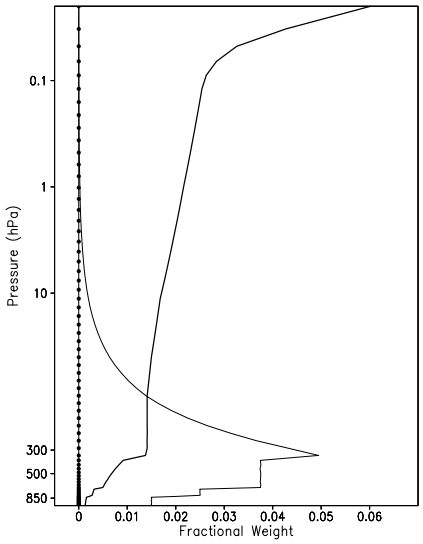


**Figure 28:** The fractional vertical weights $\Delta\sigma$ (thin curve) and $\Delta z$ (thick curve) used for calculating the ET- and EV-norms, respectively. The dotted vertical line indicates the model levels. The calculation assumes $p_s = 1000$ hPa and the model top pressure is set at 0.01 hPa. (Similar to Fig. 1 of Errico et al. 2007).

## 7.5 Job generation script

Inside the machinery of the ensemble there is a procedure called `jobgen.pl` that is used to generate job scripts to be launched within the ensemble ADAS cycle. These correspond to the various jobs submitted to the batch system while the main `atm_ens.j` driver executes. The command-line usage for `jobgen.pl` is shown in Fig. 29. Normally, users should not have to be concerned with this procedure. It should also be noted that a more general and flexible version of `jobgen.pl` is planned for and will eventually replace the one in this initial release.

```
NAME
     jobgen - Generate PBS job script

SYNOPSIS

     jobgen [...options...] jobname
                            gid
                            pbs_wallclk
                            command
                            gotodir
                            whocalled
                            file2touch
                            failedmsg

DESCRIPTION


     The following parameters are required

     jobname      name of job script to be created (will be appended
                       with j extension)
     gid          group ID job will run under
     pbs_wallclk  wall clock time for job
     command      e.g., mpirun -np $ENSGSI_NCPUS GSIsa.x
     gotodir      location to cd to (where all input files reside)
     whocalled   name of calling script (e.g., obsvr_ensemble)
     file2touch  name of file to be touched indicating a successful execution
     failedmsg   message to be issued in case of failed execution,
                  between quote marks

OPTIONS

     -egress       specify file to watch for completion of job
                     (e.g., EGRESS for AGCM)
     -expid        experiment name
     -q            specify pbs queue (e.g., datamove when archiving)
     -h            prints this usage notice

NECESSARY ENVIRONMENT

  JOBGEN_NCPUS           number of CPUS
  JOBGEN_NCPUS_PER_NODE  number of CPUS per node

OPTIONAL ENVIRONMENT

  FVROOT          location of build's bin
  ARCH            machine architecture, such as, Linux, AIX, etc
  FVHOME          location of alternative binaries
```

**Figure 29:** Job generation script command-line usage.

### 7.6 Job monitor script

Another script of very significant importance in aiding the ensemble ADAS is named `jobmonitor.csh` (see command-line usage in Appendix). As its name implies, this script has the function of monitoring each and every parallel process launched within the ensemble ADAS so that proper synchronization can take place. This is illustrated in Fig. 24, which shows four synchronization moments (boxes) during the ensemble ADAS cycle. The first makes sure the ensemble analysis (EnKF) does not start until all observers have completed their task. The second makes sure the NMC-like perturbations are available before they are needed by the re-centering procedure. The third halts the start of the ensemble of forecasts until all IAU-forcing terms are available. And lastly, the fourth makes sure post-processing of the ensemble of forecasts is only started after all forecasts have indeed completed successfully.

As discussed briefly earlier, the one important environment variable controlling the monitoring capability is `JOBMONITOR_MAXSLEEP_MIN`. This allows the user to specify the maximum amount of time (in minutes) that any particular set of parallel procedures can take to complete. For example, setting it to 60 tells the monitoring script that all observers are expected to complete within one hour; similarly, this is also the maximum time allowed for all forecasts to complete. This is a tunable parameter, and unfortunately, it is rather dependent on the ease of accessibility to the batch queue. It is possible at times to have, say, the forecasts sit in the batch queue, waiting to execute, for longer than the time allowed by `JOBMONITOR_MAXSLEEP_MIN`. Therefore, even though the main ensemble ADAS script, `amt_ens.j`, may still be running and have time to continue to run, it will quit since the processes being monitored will have run out of time. The extreme choice for `JOBMONITOR_MAXSLEEP_MIN` is setting it to match the time allowed for `amt_ens.j` to run. However, this is not a very good choice since the job may wait in vain, especially if something legitimate fails during, say, an observer or a forecast execution. Users should experiment with this environment variable and change it according to the NCCS machines status.

This page intentionally left blank.

# 8 Additional Features

## 8.1 State-space observation impact

Observation impact on the forecast can be calculated using the state-space approach of Langland and Baker (2004). This requires the availability of a model adjoint and an adjoint of the analysis system. As briefly described in Sec. 1, GMAO has two versions of atmospheric model adjoint codes available: one for its lat-lon hydrodynamics, and another for its cubed-sphere core. GMAO also has an adjoint of GSI (Trémolet 2007a; Trémolet 2008). Routine calculation of observation impact on the 24-hour forecasts, using the lat-lon model adjoint and the adjoint of GSI in its traditional 3D-Var mode, is part of our operational suite. In principle, it should be rather simple to calculate observation impact when using the *hybrid* GSI. After all, it only involves the ability to reconstruct the hybrid background error covariance matrix during the adjoint minimization. Assuming the ensemble members from the forward run have been saved, this should be a straightforward operation.

Unfortunately, this is not quite so simple due to an implementation detail. There are currently multiple options for conjugate-gradient (CG) minimization strategies in GSI. The adjoint GSI is only coded for the options within the so-called square-root-**B** preconditioning minimization strategies (i.e., standard CG, Lanczos-based CG, and Quasi-Newton). On the other hand, the hybrid capability is only implemented for the so-called **B**-preconditioning minimization strategies (i.e., double-CG and bi-CG). Essentially, the GSI adjoint requires availability of a square-root operator decomposition of each term specifying its full (climatological plus ensemble) background error covariance matrix. A square-root decomposition operator is available in the code when the background error covariance matrix is purely climatological, ($\mathbf{B} = \mathbf{B}_s$ only), but not when this matrix is hybridized with an ensemble component, $\mathbf{B}_e$. Fortunately, it is possible to approximate the adjoint when using the bi-CG minimization of El Akkraoui et al. (2013)[31]. This case only requires access to the non-decomposed background error covariance (climatological plus ensemble) operators. Since it is not yet clear to the authors how to handle sequential backward updates of the gradient in a multi-outer-loop bi-CG minimization, an approximation is made that consists of using only a single outer loop to run the adjoint bi-CG hybrid GSI. We should point out that approximating the adjoint GSI is not really a big issue. GMAO has been using an approximate adjoint in its operational observation impact suite for quite some time: the forward GSI uses the non-linear double-CG minimization with two outer-loops and a total of 250 iterations (100+150), whereas the backward GSI uses the linearized square-root-**B** preconditioned standard CG minimization with two outer-loops and a total of 200 iterations (100+100).

The way to trigger the adjoint hybrid GSI involves the following settings:

**Existing resource file.** In order for the adjoint GSI to run in hybrid mode it is necessary to make sure the namelist `HYBRID_ENSEMBLE` in the resource file `gsi_sens.rc.tmpl` is set in the exact same way as in the resource file controlling the forward hybrid GSI, namely `gsi.rc.tmpl`. Furthermore, the minimization strategy for the adjoint must be set to the bi-CG, which amounts to replacing the entry `lsqrtb = .true.` with `lbicg = .true.` in the file `gsi_sens.rc.tmpl`.

**Additional resource file.** Similarly to when replaying the hybrid ADAS from an existing ensemble, an acquire resource file is required to tell the scripts where to grab the ensemble of backgrounds. In this case, the acquire resource file is named `atmens_asens.acq`, and it must be placed in the `$FVHOME/run` directory of the experiment. The contents of this file are set in complete analogy to how the replay sets its file (see Sec. 8.3).

**Additional environment variables.** The approximation of forcing only a single outer loop minimization to take place when running the adjoint GSI is controlled by setting the environment variable `USRMITER` in the adjoint sensitivity job script `g5asens.j`. Furthermore, in analogy to what is done when replaying the hybrid ADAS, an environment variable named `HYBRIDGSI` must be set in this same

[31] The same should be feasible to implement for the default forward double-CG option of Derber and Rosati (1989), but this is not an option yet.

job script. As before, this variable should point to the working area (into which the tar-ball with backgrounds will be brought and unfolded). That is, the `g5asens.j` script should have the following extra entries:

```
setenv HYBRIDGSI $FVWORK/atmens
setenv USRMITER 1
```

The ability to run the hybrid adjoint GSI analysis assumes a forward hybrid ADAS experiment has been run and has saved the ensemble of backgrounds – by now we know this is accomplished by having the collection "ebkg" specified as part of the definition of the environment variable `ENSARCH_FIELDS` while running the forward ADAS.

## 8.2 Observation-space observation impact

In addition to observation impacts calculated with the state-space approach of Langland and Baker (2004), as briefly discussed in Sec. 8.1, observation impacts can also be calculated directly in observation space following the approach of Todling (2013). In particular, observation *impact on the analysis* can be calculated on the flysource file:

**GSI_GridComp_ensfinal.rc.tmpl.** This file is created as a copy of the file `GSI_GridComp.rc.tmpl`, and is placed under the `$ATMENSETC` directory. Edit the new file `GSI_GridComp_ensfinal.rc.tmpl` and replace the template name

```
%s.bkg.eta.%y4%m2%d2_%h2z.>>>NCSUFFIX<<<
```

with

```
%s.ana.eta.%y4%m2%d2_%h2z.>>>NCSUFFIX<<<   .
```

For now, this capability can only be exercised by strategies relying on the observer, such as the EnKF. The presence of the resource file above, together with the file `obs1gsi_member.rc`, triggers an extra call to the observer, controlled by the script `obsvr_ensfinal.csh`. This is illustrated in Fig. 24 by the double-dashed marbled box called right after the ensemble analysis controlling script. This extra observer call operates on the mean analysis. Note that the frequency of backgrounds used in the observer is set to 3 hours, whereas the frequency of analyses is set to 6 hours. That is, the observers follow a first-guess at appropriate time (FGAT; see Massart *et al.* 2010, and references therein) strategy, while the EnKF analysis is valid only at the synoptic hours, i.e., the EnKF is a filter for which the solution is valid only at a given time. To get updates for the backgrounds over the two times around the central time, the `obsvr_ensfinal.csh` script proceeds according to the 3D-Var formulation; since the increment does not evolve within the assimilation time window, updates can be obtained by simply adding the synoptic-hour increment to the two backgrounds bracketing the synoptic time. Once the update of the off-synoptic time background fields is complete, the observer can be called to produce the so-called observation-minus-analysis (OMA) residuals. Only the mean backgrounds are updated this way, thus producing ensemble mean OMA only. Observation impacts on the mean analysis can be calculated as in

$$\delta e = [\mathbf{y} - \mathbf{h}(\bar{\mathbf{x}}^a)]^T \mathbf{R}^{-1} [\mathbf{y} - \mathbf{h}(\bar{\mathbf{x}}^a)] - [\mathbf{y} - \mathbf{h}(\bar{\mathbf{x}}^b)]^T \mathbf{R}^{-1} [\mathbf{y} - \mathbf{h}(\bar{\mathbf{x}}^b)] \tag{20}$$

which can be broken up into various individual observation types. The ultimate impacts calculation is done in the program `odsstats`, when called with proper arguments. The results are placed into the Observation

Data Stream (ODS; da Silva and Redder 1995) format. To have these files stored to the archive, the collection "eoi0" should be added to the environment variable `ENSARCH_FIELDS`. We should mention that calculating the degrees-of-freedom for signal diagnostic of Lupu et al. (2011; and references therein) is also possible, requiring only a minor script change, since `odsstats` is already capable of doing this calculation.

A preliminary implementation of generating observation *impact on the mean ensemble mid-range forecasts* is presently being worked into the machinery of ensemble ADAS. More on this will appear in future releases of the software and of this document.

## 8.3 Replaying the hybrid ADAS

There are a number of reasons to have a replay capability in place. The simplest one is to have a safety net: in case analysis output files are lost or corrupted we need to be able to re-generate them by re-running the analysis for the particular cycle in question. Another, more practical reason, is the fact that not all tests and experimentations with GEOS ADAS should require a full re-generation of the ensemble. That is to say, we can see multiple instances when tests and experiments with the system have nothing to do with the ensemble and are expected to change results only mildly. These cases can rely on an already-existing ensemble of backgrounds, such as those created from an operational run, saving the user from the burden of having to run the entire ensemble-variational ADAS. Indeed, this is the recommended mode for most developers to experiment with; only when they are satisfied with their tests in non-hybrid mode do we suggest that they run a complete experiment.

Follow the steps below to run hybrid (central) ADAS experiments that simply rely on an already existing ensemble of backgrounds:

**Existing resource files.** The first thing to set properly are the resource files controlling the forward and adjoint GSI runs of your experiment, namely, `gsi.rc.tmpl` and `gsi_sens.rc.tmpl`, respectively. These files should be set to run in hybrid mode. Users should be careful to set up proper namelist parameters to ensure correct resolution of the ensemble, proper balance constraints, and other related options. Refer back to Sec. 4.2.1, for the forward hybrid GSI settings, and read on to see how to set up the hybrid adjoint GSI (Sec. 8.1).

**Additional resource file.** You must tell the scripts where to grab the existing ensemble of backgrounds from. For that, an acquire resource file name `atmens_replay.acq` must be placed in the `$FVHOME/run` directory of the experiment. All this file needs to have is a single line informing the analysis sensitivity scripts about the location of the tar-ball containing the ensemble of backgrounds. For example, the line below shows a typical content in this resource file:

```
/archive/u/USER/EXP/atmens/Y%y4/M%m2/EXP.atmens_ebkg.%y4%m2%d2_%h2z.tar
```

As usual, if the experiment name of your run is not the same as that of the run holding the ensemble, the naming can be redirected in the resource file to match your experiment name. For example, when the experiment OEXP comes from user OPS, and your (USER) experiment is named EXP, the aquire resource file should look like:

```
/archive/u/OPS/OEXP/atmens/Y%y4/M%m2/OEXP.atmens_ebkg.%y4%m2%d2_%h2z.tar => EXP.atmens_ebkg.%y4%m2%d2_%h2z.tar
```

where the line above was broken up for ease of reading – it must be a single line in the resource file. The tar-ball will be unfolded by the internal mechanisms of the analysis driver.

**Additional environment variable.** Recall that when the full ensemble-variational system is set to run coupled, the environment variable `HYBRIDGSI` is used in the main hybrid ADAS job script, `g5das.j`, to tell the scripts where to find the members of the ensemble. In that case, this variable was set to `$FVHOME/atmens`. Now, in case of replaying, this variable should simply be set to `$FVWORK/atmens`. This tells the mechanisms of the analysis script to unfold the ensemble tar-ball brought into the `$FVWORK` directory (above) in a subdirectory of itself named `atmens`. At the end of the PBS job, the ensemble members go away together with everything else in `$FVWORK`.

## 8.4 Reproducing the ensemble ADAS

It is easy to foresee situations when users will need to reproduce a cycle of the ensemble ADAS. The reasons are analogous to those behind the occasional need to reproduce (hybrid) central ADAS cycles: e.g., lost files, corrupted files, or missed output. In the ensemble case, this can only be done when the original experiment has saved its minimal set of collections to allow for reproducibility. As briefly mentioned before, the minimal set of output collections that allow for reproducibility is as follows:

**rndperts.dates** – file containing dates of NMC-like perturbations taken from the database while the original cycle ran. These files are automatically (by default) archived during the experiment (or are found under `$FVHOME/atmens` before making it to the archive).

**ebkg** – collection holding the ensemble of backgrounds.

**erst** – collection holding the ensemble of AGCM restarts.

**stat** – collection holding the ensemble statistics, of which only the ensemble mean is required for reproducibility purposes.

For example, to reproduce the ensemble ADAS analysis for 0000 UTC on 28 December 2012, the user must have the following files available:

```
yourexp.rndperts.dates.20121228_00z.txt
yourexp.atmens_ebkg.20121227_21z.tar
yourexp.atmens_erst.20121227_21z.tar
yourexp.atmens_stat.20121227_21z.tar
```

where `yourexp` represents the user experiment name. Assuming all defaults are being used, the contents of the tar-balls should be placed and organized inside the directory `$FVHOME/atmens`. The file type "rndperts.dates" should be placed in the top directory `$FVHOME/atmens`, all members from the collections "ebkg" and "erst" should be placed in subdirectories of this directory, with names identical to those in the tar-balls ("mem001", "mem002", etc), and finally, the "ensmean" subdirectory should be extracted from the collection "stat" and placed as subdirectory of `$FVHOME/atmens`.

Another thing that must be done is to create the directory defined through the environment variable `RSTSTAGE4AENS`, usually `$FVHOME/atmens/RST`, and place a copy of the file `yourexp.rst.lcv.20121227_21z.bin` inside of that. This tells the main ensemble ADAS script when the integration of the ensemble begins (remember, this file holds the valid time-stamp of the AGCM restarts).

The last thing to do is, obviously, to submit the driving script `atm_ens.j` to the batch system.

## 8.5 Experimenting with the ensemble-only ADAS

Some of us are bound to want to experiment with the GEOS ensemble-only ADAS capability. Referring back to Fig. 24, we see that one of the very last things done in the flowchart is the submission of the central hybrid ADAS job script `g5das.j`. Clearly, it does not have to be this way. And indeed, it is just as simple for the driving job script of the ensemble ADAS, `atm_ens.j`, to submit itself. This can be done by examining the definition of two environment variables set at the top part of `atm_ens.j`, namely, the variables `ENSONLY_BEG` and `ENSONLY_END`. By default, the script sets them as follows:

```
# To trigger ensemble-only set dates to anything but 0 (as yyyymmddhh)
#setenv ENSONLY_BEG 2011111221
#setenv ENSONLY_END 2011120321
setenv ENSONLY_BEG 0
```

```
setenv ENSONLY_END 0
```

As the comment says, replacing the zero-settings with actual begin and end dates tells the script not to invoke the hybrid ADAS `g5das.j` but instead to submit itself at the end of each cycle. The very first time, before cycling begins, the user will have to place the "rst.lcv" restart corresponding to the initial date and time of the cycle in the directory defined by `RSTSTAGE4AENS`, usually `$FVHOME/atmens/RST`. That is, if the dates above are used as begin and end dates of the ensemble-only cycle, then the file `expid.rst.lcv.20111112_21z.bin` must be in this directory ("expid" being the user's experiment name).

Now, before starting the ensemble-only ADAS cycle one more factor needs to be considered. We have seen in Fig. 2 that two pieces couple the ensemble and hybrid ADAS schemes: the ensemble of backgrounds that feed into the hybrid, and the central analysis that feeds into the ensemble. When running in ensemble-only mode, the non-zero setting of variables `ENSONLY_BEG` and `ENSONLY_BEG` automatically eliminates the former coupling. The latter coupling, however, must be considered carefully. If nothing else is done, the default settings of the ensemble ADAS scripts will look for the central analysis, and its corresponding satellite bias correction coefficient files, under `$FVHOME/atmens/central`. Since the central ADAS is turned off, the files will be missing, and everything will come to a halt. The solution for this is to consider a part of the main ADAS `atm_ens.j` script that was bypassed in Sec. 2 when we provided a step-by-step description for what takes place in this driver. This is the part that reads as follows:

```
if ( -e $ATMENSETC/central_ana.rc ) then
  if ( $DO_ATM_ENS ) then
   if ( ! -e $FVWORK/.DONE_MEM001_GETCENTRAL.$yyyymmddhh) then
    if(! -d $STAGE4HYBGSI ) mkdir -p $STAGE4HYBGSI
    set spool  =  "-s $FVWORK/spool"
    jobgen.pl \
           -q datamove \
           getcentral          \
           $GID                \
           $OBSVR_WALLCLOCK    \
           "acquire -v -strict -rc $ATMENSETC/central_ana.rc
                    -d $STAGE4HYBGSI $spool -ssh $anymd $anhms 060000 1" \
           $STAGE4HYBGSI       \
           $myname             \
           $FVWORK/.DONE_MEM001_GETCENTRAL.$yyyymmddhh \
           "Main job script Failed for Get Central Analysis"

           if ( -e getcentral.j ) then
              qsub  -W block = true getcentral.j
              touch .SUBMITTED
           else
              echo " $myname: Failed for Get Central Analysis, Aborting ... "
              touch $FVWORK/.FAILED
              exit(1)
           endif
    endif
  endif
endif
```

As usual, the presence of a resource file in the `$ATMENSETC` directory triggers a particular behavior in the cycle. In this case, the user must provide a resource file named `central_ana.rc` that contains the location of an existing set of analyses and bias correction files that can be used by the ensemble-only ADAS cycle. A typical example of its content is:

```
/archive/u/dao_ops/e572p5_fp/ana/Y%y4/M%m2/e572p5_fp.ana.satbang.%y4%m2%d2_%h2z.txt
     = > hy11a.ana.satbang.%y4%m2%d2_%h2z.txt
/archive/u/dao_ops/e572p5_fp/ana/Y%y4/M%m2/e572p5_fp.ana.satbias.%y4%m2%d2_%h2z.txt
     = > hy11a.ana.satbias.%y4%m2%d2_%h2z.txt
```

where here, files from the operational forward processing experiment ran with GEOS 5.7 are being fed into the user experiment (named "hy11a"; notice lines are broken up for readability purposes only). The ensemble ADAS script will still be missing the analysis file needed for re-centering. To complete the settings, the environment variable `DONORECENTER` should be set to 1 (on) in the `AtmEnsConfig.csh` configuration file. This way, no re-centering will be done, and the script will not look for the analysis file.

In summary, to run an ensemble-only ADAS:

1. Edit `atm_ens.j`, and set the begin and end date parameters `ENSONLY_BEG` and `ENSONLY_END` to desirable, non-zero, dates.

2. Consider what to do about re-centering, and define the contents of the resource file `central_ana.rc` accordingly. Depending on your choice, remember to check the environment variable `DONORECENTER` in the configuration settings of the ensemble.

3. For now, make sure the resource file `central_ana.rc` points to existing satellite bias coefficient files from another (OPS) experiment.

Ensemble purists still might dislike the fact that satellite bias correction coefficients are being brought into the ensemble-only ADAS from outside. This can be remedied when running the EnKF. The code has triggers to do the satellite bias estimation on its own and not have to rely on external information. However, we do not presently have a knob to allow the EnKF to recycle its own bias estimates; one will be added in a follow up release.

## 8.6 Spinning up the ensemble ADAS

There are times when spinning up the members of the ensemble ADAS will be necessary. This can be done rather simply by essentially running in ensemble-only mode, as just seen, but with a couple of small changes. This is the case when re-centering about an existing analysis (hybrid or not) is desirable. We can essentially follow Sec. 8.5, except that the environment variable `DONORECENTER` should be left out of `AtmEnsConfig.csh` and the resource file `central_ana.rc` should now grab existing analyses from the same place it grabs the satellite bias correction coefficient files. That is, this resource file should now be as in:

```
/archive/u/dao_ops/e572p5_fp/ana/Y%y4/M%m2/e572p5_fp.ana.eta.%y4%m2%d2_%h2z.nc4
     = > hy11a.ana.eta.%y4%m2%d2_%h2z.nc4
/archive/u/dao_ops/e572p5_fp/ana/Y%y4/M%m2/e572p5_fp.ana.satbang.%y4%m2%d2_%h2z.txt
     = > hy11a.ana.satbang.%y4%m2%d2_%h2z.txt
/archive/u/dao_ops/e572p5_fp/ana/Y%y4/M%m2/e572p5_fp.ana.satbias.%y4%m2%d2_%h2z.txt
     = > hy11a.ana.satbias.%y4%m2%d2_%h2z.txt
```

The spin up can run for as long as desired, assuming analyses and satellite bias correction files are available for the period of interest.

## 8.7 Generating climatological-like background error covariance from the ensemble

The main assumption behind hybrid assimilation is that the underlying ensemble provides a reasonable approximation to the background error covariance required by the hybrid analysis. Under this assumption, it is conceivable to think of deriving background error covariances using the same software used to derive climatological error covariances using the NMC-method. The NMC-method (see Parrish and Derber 1992 and Bannister 2008) derives a parameterization of a static (time-independent) background error covariance from differences of 24- and 48-hour forecasts. It is possible to derive a similar parameterization from the

ensemble of background fields using the same algorithm used for the NMC-method. This can be done for a single snapshot of the ensemble of backgrounds (that is, at a given time), or by collecting multiple snapshots different times. In its simplest form, a set of 6-hour background ensemble members can replace the 48-hour forecasts, with the corresponding 6-hour background ensemble mean replacing the 24-hour forecasts. When comparing results derived from the ensemble with those derived from the traditional NMC-method one must be aware of the difference in sample sizes and possible convergence issues in the algorithm due to sample size when dealing with the ensemble members. Still, one may get a pretty good idea for how an ensemble-derived background error covariance parameterization compares with the climatological background error parameterization.

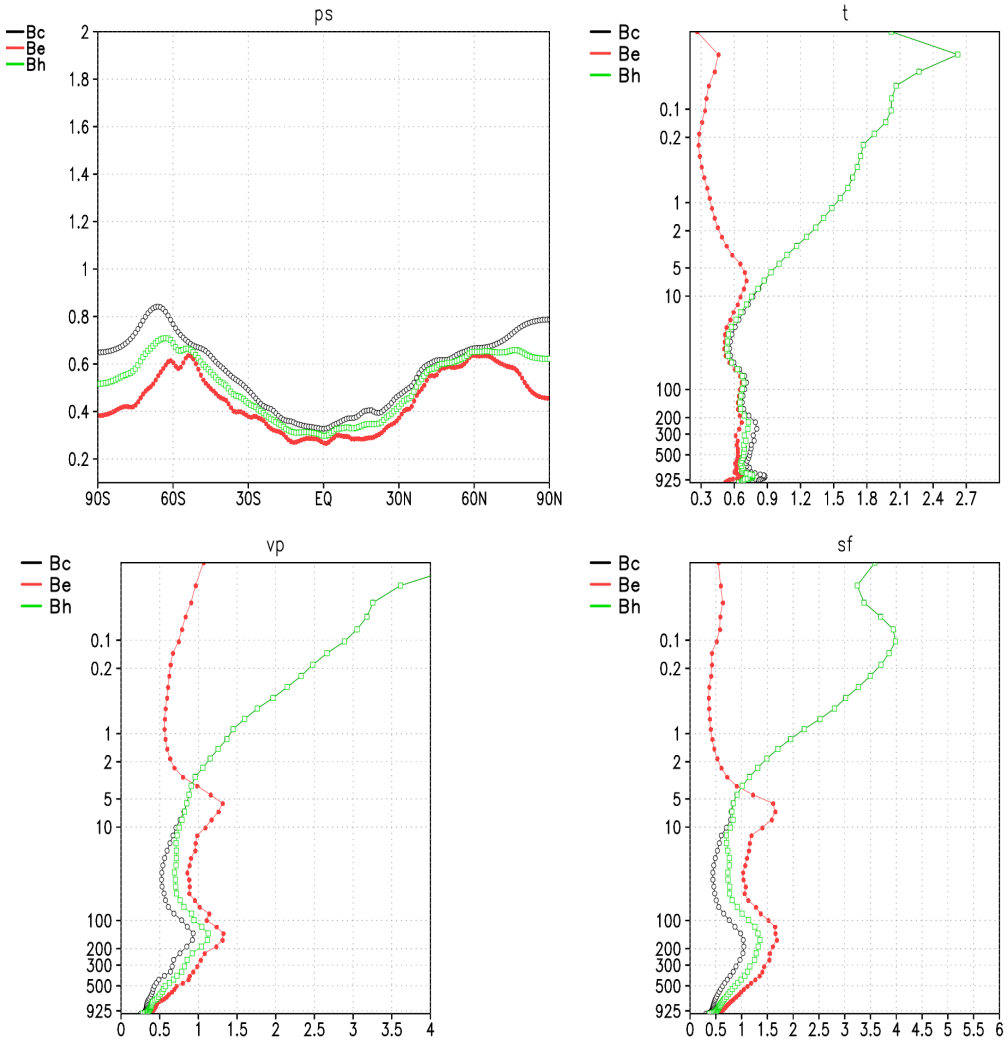


**Figure 30:** Three versions of standard deviation fields used in a parameterized background error covariance formulation: traditional NMC-method (black), ensemble-derived (red), and corresponding hybrid (green) using parameters defining the GEOS EnADAS Hybrid 3D-Var. Standard deviations are shown for the unbalanced components of surface pressure (top left), temperature (top right), velocity potential (bottom left), and total stream function (bottom right).

In the GEOS EnADAS the script named *atmens_berror.csh* is responsible for generating a "static"-like background error covariance based on the members of the ensemble. The script can be used offline (see the unit tester *ut_atmens_berror.j*), or it can be invoked online together with the central hybrid analysis. In this latter case, the generation of the ensemble-based parameterized background error covariance is called

before the hybrid analysis begins. This can be done for diagnostic purposes, or it can be used, for example, to trigger a 3D-Var like procedure in which the "static" background error covariance is actually a "covariance of the hour", replacing the climatological background error covariance of traditional 3D-Var. Discussing the validity, robustness, and consequences of using such a procedure are beyond the scope of this manuscript. The intention here is to simply make users aware of the available options in GEOS Hybrid EVADAS.

To illustrate results derived with this approach, we show in Fig. 30 estimates of standard deviation for the main meteorological fields comprising the parameterized formulation of the climatological background error covariance derived with the NMC-method using a full year of 48-minus-24-hour forecast error differences (black lines), and the equivalent quantities derived from a single instance of the 32-member ensemble from one of our experiments with GEOS Hybrid EVADAS (red lines). The variables are those forming the parameterization of (diagonal of the) background errors in GSI following Wu *et al.* (2002). The figure shows standard deviations for stream function (bottom right) and the unbalanced components of surface pressure (top left), temperature (top right) and velocity potential (bottom left). Resulting hybrid standard deviations (green lines) are also shown for each of the variables when weighted with the $\beta_c$ and $\beta_e$ coefficients shown in Fig. 25. Notice that, in this illustration, the unbalanced components of surface pressure and temperature derived from the ensemble are smaller than what comes out of the climatological background errors currently used in the GMAO GSI. The unbalanced velocity potential and total stream function are actually larger in the ensemble than in the climatological settings. A hybrid combination of these results in an overall reduction of variance in unbalanced surface pressure and temperature, along with an increase in unbalanced velocity potential and total stream function. These reductions and increases in error standard deviations that are eventually used in the hybrid GSI can be adjusted further with parameters in the GSI resource file settings ("anavinfo" table `control_vector`, see DTC GSI Documentation[32] for further details). Notice that, given our prescription of the vertically varying "beta" coefficients splitting the climatological and ensemble background error covariances, all the error standard deviations collapse to their climatological prescription above 5 hPa. Other parameters, such as horizontal and vertical correlation lengths can also be compared in a similar way. Detailed comparison and justification of the present parameter choices in the GMAO hybrid GSI are to appear elsewhere.

## 8.8 GEOS Atmospheric Ensemble Forecasting System (GAEPS)

With all the ensemble machinery presented above, enabling a script to issue ensemble forecasts from initial conditions created during the cycles of the EnADAS is rather straightforward. The GEOS atmospheric ensemble forecasting system (GAEPS) is controlled by the script name `atm_geps.j` — see lower left corner of flowchart in Fig. 24. This script sits in a subdirectory named `ageps` typically placed under `$FVHOME`. There are no required resource files for the ensemble forecasts, which means the job script can be submitted to the batch queue as long as the case to forecast has been defined. This can be done by simply creating (touching) a zero-length file with the following naming convention,

```
standalone_ageps.START_DATE_HRz+END_DATE_HRz
```

where `START_DATE_HR` and `END_DATE_HR` are to be set as in `yyyymmdd_hh`. That is, a forecast starting 2100 UTC on 1 July 1776 and ending 1200 UTC on 4 July 1776 would be handled by a file placed in `$FVHOME/ageps` and named `standalone_ageps.17760701_21z+17760704_12z`.

By default all resource files related to the configuration of model forecasts come from the `ATMENSETC` directory. However, in most cases, anyone running ensemble forecasts will want to extend the forecasts beyond the default 12 hours used for cycling the EnADAS. This entails placing a copy of the `CAP.rc.tmpl` file under the directory `$FVHOME/ageps` and adjusting it as desired. If it is desirable to have forecasts of different lengths started from different initial times `hh`, it is enough to create files `CAP_hh.rc.tmpl` under `$FVHOME/ageps`, with properly adjusted internal parameter settings related to the specific forecast

[32] Available under: https://dtcenter.org/com-GSI/users.v3.5/docs/index.php.

lengths of the various cases. In the same spirit, users can control the output (history) generated by the ensemble of forecasts by placing a file `HISTAGEPS.rc.tmpl` (or `HISTAGEPS_hh.rc.tmpl`) in the same directory. This file is a typical AGCM `HISTORY.rc` file, simply renamed and configured as needed.

Just as the script `atm_ens.j` is responsible for driving the EnADAS, the script `atm_geps.j` has internal environment variables that can be adjusted at will for expanded capability. The main variables of interest are the following:

```
setenv  RUN_AGEPS_SET  1
setenv  RUN_ENS2GCM    1
setenv  RUN_AENSFCST   1
setenv  RUN_POSTFCST   0
setenv  RUN_AENSEFS    0
setenv  RUN_AENSADFC   0
setenv  RUN_AENSVTRACK 0
setenv  RUN_ARCHATMENS 1
```

The first three, set by default, are the basic parameters required to run the GAEPS and control retrieval of ensemble initial conditions and initial ensemble analyses, calculation of IAU increment terms, and actual ensemble of forecasts, respectively. The other variables are optional and control:

**RUN_POSTFCST** — calculation of ensemble statistics for output of ensemble forecasts.

**RUN_AENSEFS** — calculation of estimate of ensemble-based forecast sensitivity vector[33].

**RUN_AENSADFC** — adjoint AGCM model integration for each ensemble member.

**RUN_AENSVTRACK** — vortex tracking program for each ensemble member.

Note that running the adjoint model for each member of the ensemble also requires: specifying an output stream in the file `HISTAGEPS.rc.tmpl` related to the member trajectory; the placement of files `CAP_apert.rc.tmpl`, `AGCM_apert.rc.tmpl`, `BACKWARD_HISTORY.rc.tmpl`, and `fvsens.ccmrun.namelist.tmpl` in the directory `$FVHOME/ageps`; users familiar with the settings of the GMAO GEOS AGCM adjoint model will recognize these files and know how to get them set up.

Behind-the-scenes script controlling GAEPS:

**atmens_prepgeps.csh** — acquires ensemble initial conditions (restarts), ensemble of inital analysis, and anything else coming from the archive or staging area

**gcm_ensemble.csh** — controls ensemble of the AGCM integrations

**post_egcm.csh** — controls post-processing of ensemble of AGCM outputs

**atmens_efsens.csh** — controls ensemble of AGCM adjoint-based forecast sensitivity generation

**atmens_vtrack.csh** — controls vortex tracking programs

Note that many of the scripts listed above are the same as those driving the EnADAS application.

[33] This feature presently follows Ancell and Hakim (2007) and is still under testing and evaluation.

## 8.9 Vortex track for ensemble members

Either within the context of the GAEPS or within the context of the EnADAS, when this latter extends its model integrations beyond the typical 12-hour integrations required for the assimilation cycling, the scripts controlling the ensemble can be told to invoke the vortex tracking scripts and programs. The core of this program comes from an adaptation of NCEP program of Marchok (2010) (see also Trahan and Sparling 2012) to track maximum relative vorticity along with minimal wind magnitude, geopotential height, and sea-level pressure in a specified area around a tropical cyclone. This program requires data from the so-called TC-Vitals[34] information made available in real-time from the NOAA distribution site. Similar to other applications, it is the presence of certain resource files that triggers this capability in the GEOS ensemble framework: when using this feature within the EnADAS cycle the following files should be placed under `$ATMENSETC`: `vtrack.ctl.tmpl`, *vtrack.rc*, and `vtx.ctl.tmpl`; when trying to trigger this feature within GAEPS these files should be placed under `$ATMENSAGEPS`. Specification of the options `-vtxrlc` of `setup_atmens.pl` seen in Sec. 3 automatically places these files in the proper locations.

An illustration of tracks derived for hurricane Katrina in 2005 is given in Fig. 31. In this case, model integrations from a cycling EnADAS were extended to run up to three days. The vortex tracking programs were triggered by the presence of the resource files just mentioned in `$ATMENSETC`. The tracking program was automatically run for each member of the ensemble and the track information of the members of the ensemble were stored in the output class `evtk` (see Sec. 4.1.6). The particular example here comes from an experiment performed with the ADAS at its current full resolution (near-real-time column of Table 1). As we have seen above, this is when the ensemble members are at the resolution of the single realization of MERRA-2 analyses. Again for illustration purposes, we ran forecasts from MERRA-2 analysis, tracking Katrina in the same way as the storm is tracked in the ensemble members' forecasts and in the full resolution deterministic forecasts. The tracks from the high resolution (C720) forecasts (solid red curves) are shown together with tracks for the C180 forecasts from MERRA-2 (solid green curves), and the C180 forecasts for each of the 32 ensemble members (solid grey curves) — the ensemble mean of the tracks appears as the yellow solid curves. The TC-vitals positions are displayed as fat dots colored to show the intensity of the storm (warm colors indicate high intensity). Each panel displays tracks calculated from forecasts issued from different initial conditions, at consecutive 0000 UTC times from 24 to 29 August 2005. Just a visual inspection is enough to see that the high resolution (C720) forecasts remain considerably more on track than the coarser resolution forecasts from either MERRA-2 or the ensemble. Though there are not enough cases here to make an accurate assessment, the mean ensemble track *seems* to be as close to the TC-Vitals observations as the tracks of the forecasts from MERRA-2. Interestingly, the background fields of MERRA-2 are relocated [35] before being fed to the GSI analysis, whereas no relocation is applied to the background members of GEOS EnADAS.[36]

Behind-the-scenes script controlling track calculation:

**atmens_vtrack.csh** — controls calculation of storm track from ensemble of forecasts

## 8.10 GEOS Atmospheric Ensemble Forecast Sensitivity and Observation Impact (GAEFSOI)

An adjoint-based forecast sensitivity and observation impact (FSOI) tool has been available in the GMAO ADAS for quite some time. The tool, initially implemented based on the adjoint model of Giering *et al.* (2017), uses the GSI adjoint of Trémolet (2007a) and has supported a variety of studies (e.g., Errico *et al.* 2007, Daescu and Todling 2009, Daescu and Todling 2010, Gelaro *et al.* 2010, Todling 2013). An upgraded version of the Giering *et al.* (2017) adjoint that uses the cubed-sphere hydrodynamics of the latest version of GEOS AGCM (e.g, Holdaway *et al.* 2014) supports the current GEOS FP observation impact tool. All

[34] NCEP/EMC: Available online at http://www.emc.ncep.noaa.gov/mmb/ data_processing/tcvitals_description.htm.

[35] Tropical strom relocation in GEOS ADAS follows http://iwintest.nws.noaa.gov/om/tpb/472.pdf.

[36] This differs from the approach followed by NCEP, where the background members are relocated before the EnSRF analysis.

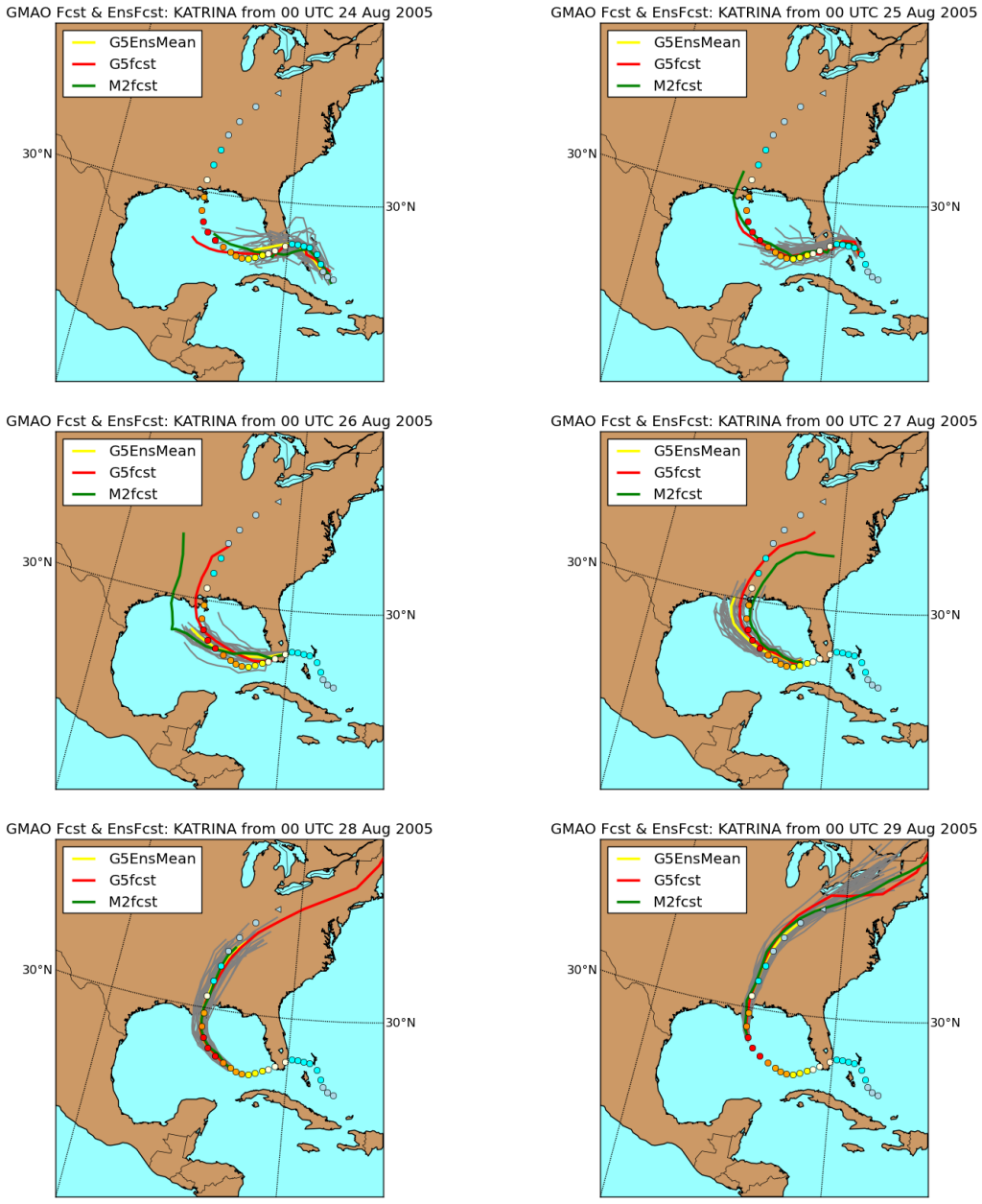


**Figure 31:** Illustration of GEOS Atmospheric Ensemble Prediction System and storm tracking capability. Three-hour tracks of 0000 UTC forecasts for Katrina from 24 to 29 August 2005 are shown. Forecasts from high resolution (near-real-time configuration of Tab. 1) cycled hybrid deterministic analysis are shown by the solid red curves; forecasts from corresponding 32-member C180 ensemble analyses are shown by the solid grey curves, with ensemble mean track shown by the solid yellow curves. Also displayed for the purposes of comparison are forecasts issued from the C180 analysis of MERRA-2 (solid green curves). Actual observed tracks from TC-Vitals database are shown by the fat dots with coloring referring to the storm's intensity (the warmer the colors, the more intense the storm).

of the GMAO work published to date on observation impact employs 3D-Var for the underlying analysis — no refereed publications have yet appeared based on the use of a hybrid analysis [37]. In point of fact, in upgrading GEOS ADAS to Hybrid 3D-Var, and more recently to Hybrid 4D-EnVar, no changes other than

[37] This is not to say that plenty of work has not been done with the hybrid analysis option, e.g., see Todling et al. 2017: https://www.jcsda.noaa.gov/documents/meetings/wkshp2017/dayThree/Todling_JCSDAwkshp2017.pdf.

trivial parameter settings had to be implemented to the observation impact software. The observation impact tool is essentially blind to the underlying analysis strategy.

Under GEOS EnADAS, availability of an ensemble of model predictions opens the door to the possibility of implementing an ensemble-based forecast sensitivity and observation impact (EFSOI) strategy following Liu and Kalnay (2008) and Li *et al.* (2010). The basic work to enable EFSOI in the EnKF software has been done by Ota *et al.* (2013). More recently Groff (2017) merged this initial implementation into the current version of the EnKF software. The first author of this tech memo and Fabio L. R. Diniz (while visiting GMAO from CPTEC, Brazil) made necessary adjustments to the EnKF/EnSRF software so it properly accounts for the observations *truly* used by the analysis (along the lines of the considerations in Sec. 1.2.1) and implemented an EFSOI capability for GMAO EnADAS.

The EFSOI calculates the following scalar corresponding to the forecast error reduction, $\delta e$, associated with the assimilation of observations between two consecutive analysis cycles:

$$\delta e = \mathbf{d}_{ob}^T \mathbf{R}^{-1} \mathbf{H} \mathbf{X}^a (\mathbf{X}^f)^T \mathbf{E} (\varepsilon^a + \varepsilon^b) \tag{21}$$

where $\mathbf{d}_{ob}$, $\mathbf{R}$, and $\mathbf{H}$ correspond, respectively, to the OMB residual vector, observation error covariance matrix, and the Jacobian of the observation operator, as introduced in earlier sections; the matrix $\mathbf{E}$ defines a metric to evaluate the impacts and typically corresponds to the tropospheric form (16) of the linearized energy operator described in Sec. 7.4; and the matrices $\mathbf{X}^a$ and $\mathbf{X}^f$ are 3D-equivalents to the ensemble perturbation matrices introduced in (5), but now corresponding to analysis and forecast ensemble perturbations. The vectors $\varepsilon^a$ and $\varepsilon^b$ are ensemble mean forecast errors, typically verified against the ensemble mean analysis, and corresponding to two forecasts started from analysis at two consecutive times, that is, six hours apart from each other. It is typical to derive observation impacts on the 24-hour forecasts, when then these error vectors are calculated for the 24- and 30-hour ensemble mean forecasts — in this case, $\mathbf{X}^f$ corresponds to a matrix of 24-hour ensemble forecast perturbations.

In order to be able to apply the expression (21), one must have access to both matrices $\mathbf{X}^a$ and $\mathbf{X}^f$. The first one comes from saving the non-inflated ensemble analyses of the EnSRF during the (forward) regular ensemble assimilation cycle. This can be done by simply adding the following variable setting to the EnSRF setup namelist — in file `atmos_enkf.nml.tmpl`:

```
fso_cycling = .true. .
```

The above triggers storage of the required analyses in the data class `eniana` seen in Sec. 4.1.6. Obtaining the 24-hour forecast perturbation matrix $\mathbf{X}^f$ and the 24- and 30-hour ensemble mean forecast errors $\varepsilon^a$ and $\varepsilon^b$ can be accomplished by simply running 24- and 30-hour GAEPS's as discussed above.

With all required quantities available, the actual EFSOI calculation (21) involves using the same EnSRF software but choosing slightly different namelist options than those chosen to run it in forward mode. To distinguish the forward and "backward" EnSRF runs, the file holding the options for the EFSOI settings is named `atmos_enkf_sens.nml.tmpl`. For the most part, the contents of this file are identical to the contents of the file `atmos_enkf.nml.tmpl` controlling the forward case except for the inclusion of the following additional parameter settings:

```
fso_flag = .true.,
fso_cycling = .false.,
fso_cnt_usedob_only=.true.,
fso_have_ferr=.true.,
adrate=0.75,
evalft=24,
```

where `fso_flag` simply triggers the EFSOI calculation in the EnSRF; `fso_cycling` ensures that the ensemble of non-inflated analyses will not be written out in this case; `fso_cnt_usedob_only` forces

proper accountability of used observations as explained in Sec. 1.2.1; `fso_have_ferr` informs the software that it is reading pre-computed (offline-calculated) total normalized ensemble mean forecast errors (that is: $\mathbf{E}(\varepsilon^a + \varepsilon^b)$) as opposed to reading the 24- and 30-hour ensemble mean forecasts and the verification and then performing the normalization and addition of errors internally in the EnSRF software[38]; `adrate` specifies the advection rate of covariance localization scales (see Ota *et al.* 2013 for details); and `evalft` specifies the length (in hours) of the forecast associated with EFSOI.

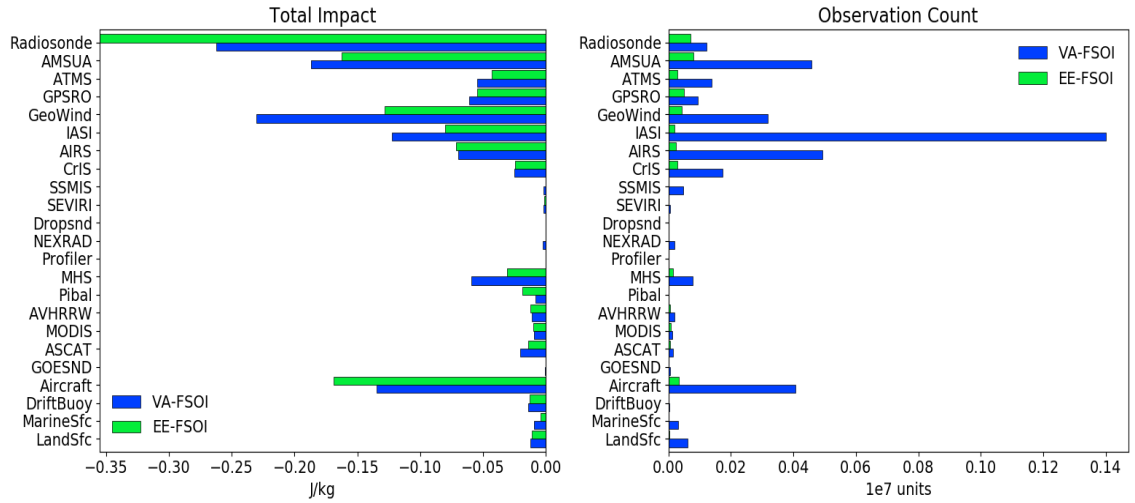


**Figure 32:** Left panel: observation impacts on the 12-hour forecasts obtained with the traditional adjoint-based tool (blue bars; labeled VA-FSOI), and equivalent impacts derived with the ensemble-based approach described in this section (green bars; labeled EE-FSOI). Right panel: observation count for observations actually used in the GSI analysis (blue bars) and in the EnSRF ensemble analysis (green bars). Results are decomposed into main instruments participating in the GEOS atmospheric analysis. (Results were obtained in collaboration with F. L. R. Diniz from CPTEC, Brazil.)

With the available ensemble forecasts and the parameter settings for the "backward" EnSRF adjusted as above, all the user needs to do is touch a file in the directory controlling analysis sensitivity experiments, typically `$FVHOME/asens`, and launch the main driver script for this feature, namely, `atm_efsoi.j`. An example of this procedure is as follows:

```
cd $FVHOME/asens
touch standalone_aensosens.20161210_00z
qsub atm_efsoi.j
```

where here, ensemble-based observation impacts will be calculated for the analysis on 0000 UTC of 10 December 2016. Although automating this latter part is relatively simple, no attempt has yet been made to do so since the full evaluation of EFSOI and comparison with FSOI is still underway, and it is not yet known how we will proceed with respect to the implementation of this feature in the GMAO FP system. Indeed, early results indicate that unless the EnSRF settings are changed so the filter assimilates as many observations as possible (and as many observations as the hybrid GSI), results from EFSOI are not representative of how observations actually impact the forecasts. This is illustrated in Fig. 32 where observation impacts on the 12-hour forecasts are derived from both the traditional adjoint-based approach (blue bars) and from the ensemble-based approach described here (green bars). Though some differences are noticeable when comparing the impact results from these two approaches, overall, one might be inclined to conclude that the

[38] Doing the calculation online is the apparent preferred mode at NCEP; GMAO does the error calculation offline for consistency with how its adjoint-based FSOI calculations are performed

results are acceptably comparable. This, however, is only until we realize that the number of observations being used by the two underlying analyses, namely, GSI and EnSRF, are dramatically different (right panel). To make a fair comparison one would have to get the EnSRF to use a number of observations comparable to what GSI uses. At the time of this writing, a study is ongoing to make as consistent a comparison as possible; results of this study will appear elsewhere.

Users should be aware that EFSOI is a new, rather fresh, feature of GEOS Hybrid ADAS and that further adjustments to software and supporting scripts are still expected to take place.

Behind-the-scenes scripts controlling GEOS EFSOI application:

**atmens_prepobsens.csh** — acquires non-inflated ensemble analyses, ensemble forecasts, verifying analysis, and anything else required for this application

**atmos_enkf.csh** — controls EnSRF ensemble analysis (forward and "backward" cases)

**post_efso.csh** — controls post-processing of EFSOI output from EnSRF

## 8.11 Setting up reanalysis-like experiments

One of the main components of the GMAO's mission is to contribute to reanalysis efforts. Although this document is largely focused on the GMAO near-real-time, forward processing system, much of what is discussed here applies just as well when the system is configured for reanalysis applications. The main differences between experiments aimed at studies directed toward our reanalysis efforts as opposed to experiments directed toward investigations related to our near-real-time application are (i) the resolution configuration of the ADAS, and (ii) the configuration of the observing system component. When it comes to (i), configuring the system covered in this document to run at, say, the MERRA-2 C180 resolution (see Gelaro and coauthors 2017) is relatively simple as far as the central ADAS is concerned; most users of GEOS ADAS are quite familiar with this aspect. Under the hybrid option, there is the additional need for having to configure the resolution of the EnADAS for the particular reanalysis-like experiment. This, however, requires no extra information than what is provided in Sec. 4.1; indeed, resolution is trivially set when running `atmens_setup.pl` (see Sec. 3). The only topic of concern relates to (ii), the configuration of the observing system. The GEOS ADAS point to a very specific observation database when it comes to reanalysis experiments. Details of the MERRA-2 databased appear in McCarty *et al.* (2016).

Beside the difference in the observing systems database used for near-real-time (FP) and reanalysis applications, there is also a difference in the way the conventional observing system is treated in these cases. In the FP system GMAO uses the so-called NCEP "prepbufr" files[39], which provides a quality controlled version of conventional observations to be ingested in the GSI observer. This part of the database is made available to GMAO in real time and includes observations quality controlled by NCEP. In other words, the GMAO FP system does not quality-control observations offline using its own backgrounds and the updated version of Woollen *et al.* 1994 quality control, but instead simply directly uses the "prepbufr" observations in the central and ensemble analyses. Conversely, in reanalysis mode, we choose to perform the offline quality control step to handle the raw observations, combined with our own backgrounds, before giving the data to the GSI observer.

Instead of getting too much into the details of the observing systems and particular quality control issues, we provide here an example of observing systems settings for both FP and reanalysis applications. As briefly mentioned in Sec. 4.1.2, in GEOS ADAS, the observing system is controlled by the environment variable `OBSCLASS`. This variable is a conglomerate of so-called observing system *classes*, which are typically associated with differing observing instruments, or sometimes sets of instruments such as the class `ncep_prep_bufr` referring to the conventional observations from NCEP. An experiment set to run like the FP system might have this environment variable set as follows:

```
setenv OBSCLASS "ncep_prep_bufr,ncep_1bamua_bufr,ncep_1bamub_bufr,
```

[39] Information on these files can be found under http://www.emc.ncep.noaa.gov/mmb/data_processing/prepbufr.doc/document.htm.

```
                    ncep_1bhrs2_bufr,ncep_1bhrs3_bufr,ncep_1bmsu_bufr,
                    disc_airs_bufr,disc_amsua_bufr,ncep_mhs_bufr,
                    ncep_1bhrs4_bufr,ncep_goesfv_bufr,ncep_mtiasi_bufr,
                    ncep_gpsro_bufr,ncep_aura_omi_bufr,ncep_satwnd_bufr,
                    ncep_atms_bufr,ncep_sevcsr_bufr,ncep_cris_bufr,
                    ncep_ssmis_bufr,mls_nrt_nc,rscat_bufr,ncep_avcsam_bufr,
                    ncep_avcspm_bufr,ncep_tcvitals,gmao_gmi_bufr"
```

with the various observing classes being explicitly defined in a resource file named `obsys.rc`. It is not necessary to go over each of the classes included above, but many are very much self evident. For example, `disc_airs_bufr` controls use of observations from the Atmospheric Infrared Sounder (AIRS) instrument on the AQUA satellite; `ncep_sevcsr_bufr` controls use of observations from Spinning Enhanced Visible and InfraRed Imager (SEVIRI) instruments; and so on.

Alternatively, a reanalysis-like experiment would set `OBSCLASS` in the following way:

```
setenv OBSCLASS0 "merra2_cdas_pre-qc_bufr,merra2_avhrrwnd_pre-qc_bufr,
                  merra2_repro_ers2_pre-qc_bufr,
                  merra2_qscat_jpl_pre-qc_bufr,
                  merra2_wspd_pre-qc_bufr,merra2_prof_pre-qc_bufr"

setenv OBSCLASS1 "merra2_ncep_tcvitals,merra2_gprofp13_bufr,
                  merra2_gprofp14_bufr,merra2_tmil_bufr,merra2_tmio_bufr,
                  merra2_gpsro_bufr,merra2_goesnd_prep_bufr,
                  merra2_1bamua_bufr,merra2_1bamub_bufr,merra2_1bhrs2_bufr,
                  merra2_1bhrs3_bufr,merra2_1bmsu_bufr,merra2_1bssu_bufr,
                  merra2_eosairs_bufr,merra2_eosamsua_bufr,
                  merra2_ssmit13_bufr,merra2_ssmit14_bufr,
                  merra2_ssmit15_bufr,merra2_aura_mlsoz_bufr,
                  merra2_aura_omieff_nc,merra2_gmao_mlst_bufr"

setenv OBSCLASS "$OBSCLASS0,$OBSCLASS1"
```

In this case, the general `OBSCLASS` is broken up into two components. The first component, defined by `OBSCLASS0`, relates to various subcomponents of the conventional observing network (radiosondes, aircrafts, bouys, etc). These are the classes that are to be processed by the offline quality control package based on the GEOS ADAS background fields; notice how all the class names include an inner name `pre-qc` indicating these components correspond to the pre-quality control (raw) observations. The second set of observation classes defined by `OBSCLASS1` includes all other observing systems not handled by the offline quality control (i.e., quality controlled by the GSI observer directly). When the quality control processes the `pre-qc` classes it puts all the data together into a GMAO-generated "prepbufr" file corresponding to an associated class named `gmao_prep_bufr`.

When the EnADAS is used in the context of the GEOS Hybrid ADAS, and a reanalysis-like experimental scenario is employed, the ensemble controlling scripts are not required to re-run the offline quality control from the raw observations. In this case, the EnADAS is instructed to use the quality-controlled observation class `gmao_prep_bufr`. That is, the ensemble observers use observations quality-controlled by the central ADAS. This is specified by having the following line appended to the ensemble configuration file `AtmEnsConfig.csh` (see Sec. 4.1.2):

```
setenv OBSCLASS "gmao_prep_bufr,$OBSCLASS1"
```

which simply overwrites the subcomponents of the observation class corresponding to the raw observations with that corresponding to the quality-controlled data.

Of relevance here too are the settings related to the observing classes in the GSI configuration file, `gsi.rc.tmpl`. Typically, when experimenting in an FP-like scenario, the `OBS_INPUT` namelist settings found in this file look something like the following:

```
OBS_INPUT::
!  dfile          dtype       dplat       dsis                dval    dthin  dsfcalc  obsclass
   prepbufr       ps          null        ps                  0.0     0      0        ncep_prep_bufr
   prepbufr       t           null        t                   0.0     0      0        ncep_prep_bufr
   prepbufr       q           null        q                   0.0     0      0        ncep_prep_bufr
   prepbufr       uv          null        uv                  0.0     0      0        ncep_prep_bufr
   prepbufr       spd         null        spd                 0.0     0      0        ncep_prep_bufr
   radarbufr      rw          null        rw                  0.0     0      0        ncep_prep_bufr
   prepbufr       dw          null        dw                  0.0     0      0        ncep_prep_bufr
   prepbufr       pw          null        pw                  0.0     0      0        ncep_prep_bufr
   preprscat      uv          null        uv                  0.0     0      0        rscat_bufr
   gpsrobufr      gps_bnd     null        gps                 0.0     0      0        ncep_gpsro_bufr
   ssmirrbufr     pcp_ssmi    dmsp        pcp_ssmi            0.0     0      0        ncep_spssmi_bufr
   tmirrbufr      pcp_tmi     trmm        pcp_tmi             0.0     0      0        ncep_sptrmm_bufr
   sbuvbufr       sbuv2       n16         sbuv8_n16           0.0     0      0        ncep_osbuv_bufr
   sbuvbufr       sbuv2       n17         sbuv8_n17           0.0     0      0        ncep_osbuv_bufr
   .
   .
   .
::
```

where only a small part of the namelist is displayed here for simplicity. Notice the last column of the table in the namelist refers to the observation class[40]. These names are directly related to the settings of the environment variable `OBSCLASS` as discussed above. In a reanalysis-like scenario, the observation classes in the `gsi.rc.tmpl` must change accordingly, as for example in what follows:

```
OBS_INPUT::
!  dfile      dtype        dplat       dsis                dval    dthin  dsfcalc  obclass
!Conventional
   prepbufr    ps          null        ps                  0.0      0     0        gmao_prep_bufr
   prepbufr    t           null        t                   0.0      0     0        gmao_prep_bufr
   prepbufr    q           null        q                   0.0      0     0        gmao_prep_bufr
   prepbufr    uv          null        uv                  0.0      0     0        gmao_prep_bufr
   prepbufr    spd         null        spd                 0.0      0     0        gmao_prep_bufr
   prepbufr    sst         null        sst                 0.0      0     0        gmao_prep_bufr
   prepbufr    pw          null        pw                  0.0      0     0        gmao_prep_bufr
   satwndbufr  uv          null        uv                  0.0      0     0        merra2_satwnd_bufr
   satwndavhr  uv          null        uv                  0.0      0     0        merra2_avhrr_satwnd_bufr
   tcvitl      tcp         null        tcp                 0.0      0     0        merra2_ncep_tcvitals
   mlstbufr    t           aura        t                   0.0      0     0        merra2_gmao_mlst_bufr
   oscatbufr   uv          null        uv                  0.0      0     0        merra2_oscat_bufr
!Infrared
   hirs2bufr   hirs2       tirosn      hirs2_tirosn        0.0      1     0        merra2_1bhrs2_bufr
   hirs2bufr   hirs2       n06         hirs2_n06           0.0      1     0        merra2_1bhrs2_bufr
   hirs2bufr   hirs2       n07         hirs2_n07           0.0      1     0        merra2_1bhrs2_bufr
   .
   .
   .
::
```

The important thing to notice is that, in this case, the conventional observations ("prepbufr") come from the GMAO quality-controlled class `gmao_prep_bufr`; all other classes are the MERRA-2 corresponding classes.

Finally, to allow the ensemble observers to use the "prepbufr" file created by the central ADAS it is necessary to define the data-class `gmao_prep_bufr` in the observation database, i.e., in `obsys.rc`. The following provides an example of what needs to appear in this resource file:

```
BEGIN gmao_prep_bufr => USREXPID.prepbufr.%y4%m2%d2.t%h2z.bin
  20050625_00z-21001231_18z 240000 FVHOME/atmens/central/USREXPID.prepbufr.%y4%m2%d2.t%h2z.blk
END
```

Users slightly familiar with GEOS ADAS will know that each class of the database defines the location where files are to be found. In the case here, the default is for the central ADAS to place a copy of the "prepbufr" file under `$FVHOME/atmens/central`. The combination of the redefined `OBSCLASS` appearing in `AtmEnsConfig.csh` and the definition of the class `gmao_prep_bufr` in `obsys.rc` is each to allow the ensemble observers to access the "prepbufr" observations quality-controlled by the central ADAS.

[40]This column is not standard in the official GSI release from DTC; this is a GMAO-specific construct used by the ADAS controlling scripts, and never relating to the actual Fortran code.

# 9 Conventions

Each process launched by GEOS ensemble ADAS indicates its successful termination by touching a properly named hidden file in the work-directory of the ongoing execution. An example of the hidden files present near completion of integration of a 3-member ensemble experiment is given below.

```
.DONE_ENSMEAN_obsvr_ensemble.csh.2012032600
.DONE_MEM001_obsvr_ensemble.csh.2012032600
.DONE_MEM002_obsvr_ensemble.csh.2012032600
.DONE_MEM003_obsvr_ensemble.csh.2012032600
.DONE_acquire_ensperts.csh.2012032600
.DONE_MEM001_ACQUIRE_ENSPERTS.2012032600
.DONE_obsvr_ensemble.csh.2012032600
.DONE_PERTMEAN.2012032600
.DONE_ENKFX_atmos_enkf.csh.2012032600
.DONE_atmos_enkf.csh.2012032600
.DONE_atmos_eana.csh.2012032600
.DONE_MEM001_setperts.csh.2012032600
.DONE_MEM001_PERTDIFF.2012032600
.DONE_MEM002_PERTDIFF.2012032600
.DONE_MEM003_PERTDIFF.2012032600
.DONE_MEM002_atmens_recenter.csh.2012032600
.DONE_MEM001_atmens_recenter.csh.2012032600
.DONE_MEM003_atmens_recenter.csh.2012032600
.DONE_atmens_recenter.csh_ana.eta_ana.eta.2012032600
.DONE_post_eana.csh.2012032600
.DONE_MEM001_atmos_ens2gcm.csh.2012032600
.DONE_MEM002_atmos_ens2gcm.csh.2012032600
.DONE_MEM003_atmos_ens2gcm.csh.2012032600
.DONE_atmos_ens2gcm.csh.2012032612
.DONE_MEM001_gcm_ensemble.csh.2012032609
.DONE_MEM002_gcm_ensemble.csh.2012032609
.DONE_MEM003_gcm_ensemble.csh.2012032609
.DONE_ENSFCST
```

This more or less tells the sequence of events in the run. For example, the run begins by launching the acquire of NMC-like perturbations to be used as additive inflations. The observers run concurrently to this. The mean observer runs first and indicates its successful termination by touching the zero-length file

```
.DONE_ENSMEAN_obsvr_ensemble.csh.2012032600
```

Following the mean observer, each member observer runs concurrently to each other. As they terminate successfully, they touch corresponding hidden files indicating the ensemble member number, e.g., the third member touches

```
.DONE_MEM003_obsvr_ensemble.csh.2012032600
```

For consistency, a time stamp is also appended to the touched hidden file. When all observers are finished, the hidden file ".DONE_obsvr_ensemble.csh.2012032600", with the calling program name, appears. When all NMC-like perturbations have been retrieved, the file

```
.DONE_MEM001_ACQUIRE_ENSPERTS.2012032600
```

indicates their availability to the experiment. The file ".DONE_PERTMEAN.2012032600" indicates the successful removal of the mean from the NMC-like perturbations. Analogously, when all processes related to the ensemble analysis complete, the file ".DONE_atmos_eana.csh.2012032600" appears. The availability of both the mean-free NMC-like perturbations for additive inflation and the ensemble analyses allows the re-centering to take place, with successful termination indicated by the file ".DONE_atmens_recenter.csh_ana.eta_ana.eta.2012032600". This is followed by the completion of the post-analysis, which involves collection of statistics from the ensemble; completion is indicated by

```
.DONE_post_eana.csh.2012032600
```

Once re-centering is finished, the IAU increments are created for each member by the procedure connecting the ensemble analyses with the AGCM,

```
.DONE_atmos_ens2gcm.csh.2012032612
```

Finally, the ensemble of backgrounds for the next cycle is created by running instances of the AGCM for each member. Each member of the forecast indicates its own successful termination, and when all are done, the file ".DONE_ENSFCST" indicates the successful generation of all new members. At this point, the main job script will take care of submitting the central ADAS cycling script and launching the archiving job to save requested output from the ensemble ADAS. This brief description is sequential but by now the reader will be familiar with the parallelism exploited within this process.

The namings of the hidden files are standardized. That is, their names follow the template:

```
.DONE_CALLING_PROGRAM.YYYYMMDDHH
```

where `CALLING_PROGRAM` and `YYYYMMDDHH` specify the calling program and the current date and time of the cycle (or background). The other convention is that the hidden files must, in the majority of cases, be created at the level of the `ENSWORK` directory. There are some exceptions, but those are outside the scope of the present document. *Users wanting to implement extra features must follow the conventions*.

# 10 Handling Crashes

All systems are due to sporadic failures; the ensemble component of the ADAS is no different. Fortunately, most crashes simply require re-submitting the driving script. For now, we only discuss issues related to reviving a crashed execution when the scheduler is not running, that is to say, when the `g5das.j` submits the driving ensemble script `atm_ens.j`, which in turn submits `g5das.j`.

When various attempts to revive a crashed cycle fails, the last-resort thing to do is simply to remove the `$ENSWORK` directory and re-submit the ensemble job control script, `atm_ens.j` from the start. Before doing so, make sure the original ensemble for the present cycle of interest remains intact.

This page intentionally left blank.

# 11 Frequently Asked Questions

1. *What's the easiest way to create a set of ensemble members from scratch?*
After multiple trials, with multiple schemes, we have come to find that a viable ensemble can be created by simply reproducing the initial conditions at the time of interest to create as many members as desired. This applies to both the AGCM restart files and to the background fields. For example, follow these steps:

   **Inside $FVHOME/atmens:** Create as many directories for the members as needed. For example, for 32 members, create the directory `$FVHOME/atmens` and do:

```
cd $FVHOME/atmens
@ n  =  0
@ nmem  =  32
while ($n < $nmem)
   @ n++
   set memtag   =  `echo $ic |awk '{printf "%03d", $1}'`
   mkdir mem$memtag
end
```

   **Model restarts:** Generate model restarts for desired ensemble resolution[41] and place them in, say, a subdirectory of `$FVHOME/atmens` named `rst`. Now, link the binary restarts of the AGCM into the member directories by doing:

```
cd $FVHOME/atmens
foreach dir (`ls -d mem0*`)
   cd $dir
   ln -sf ../rst/*.bin .
   cd -
end
```

   **Background files:** Convert background files from start-up experiment (say, operational run) to desired ensemble resolution. For example, assuming you have the so-called "bkg.eta" files under `$FVHOME/atmens/rst` as described above for the binary restarts, you can convert each of the background restarts by doing:

```
$DASBIN/dyn2dyn.x -g5 -res c \
-o hy11a.bkg.eta.20121226_03z.nc4 \
e572p5_fp.bkg09_eta_rst.20121226_03z.nc4
```

   where `$DASBIN` is the location of an ADAS build `bin` directory, and the example takes a background file from the GEOS-5.7 operational series to 1-degree resolution. You should not need to worry about the resolution of the surface background files - the observers take care of resolution discrepancies on the fly. After converting the backgrounds, you can then do a similar "foreach" loop as shown above to link all "bkg.eta" and "bkg.sfc" from `$FVHOME/atmens/rst` into the directories of the individual members.

   **Ensemble mean files:** Since all background files are identical, you can use the same set of resolution-converted background files to feed a directory named `ensmean`, under `$FVHOME/atmens`.

   You can check the usage of the scripts `gen_ensbkg.csh` and `gen_ensrst.csh` since they basically do what has just been described (see Appendix).

[41] Use the `regrid.pl` utility to convert files from GMAO operational runs.

2. *What do I submit next?*
   Sometimes, when a cycling job stops, one might have difficulty knowing what to submit next, namely either the central ADAS script `g5das.j` or the ensemble ADAS script `atm_ens.j`. The first thing to check is the run directory to try and determine which PBS output file has been written out last. If that does not help, the next thing to check is the presence of the directory defined by `RSTSTAGE4AENS`[42]; if it is present, chances are the ensemble ADAS was running when the job stopped, and what likely needs to be submitted next is the ensemble ADAS script. Another possibility is to check for the presence of work area directory `$ENSWORK`. If this directory is present, then most certainly the ensemble ADAS was running when the job stopped and `atm_ens.j` should be re-submitted.

3. *How do I add new output streams to the ensemble of AGCMs?*
   The file `$ATMENSETC/HISTAENS.rc.tmpl` controls the history of each ensemble member dealt with by the AGCM. In principle, nothing forbids this history to be as complex and complete as that of the AGCM when running the central ADAS, that is, the `HISTORY.rc.tmpl` file under `$FVHOME/run`. However, each data stream added to the ensemble members results in increased wall-clock time when running each member of the ensemble. Therefore, though the capability is there, the flexibility to get the extra output stream saved (archived) is somewhat hidden. Not only will the user need to edit `$ATMENSETC/HISTAENS.rc.tmpl` but he or she will also need to edit the update and archiving scripts (`update_ens.csh` and `atmens_arch.csh`) to add logic for handling the new data stream. Eventually, as long as the same mechanism of handling archiving output is used, the environment variable `ENSARCH_FIELDS` will need to be edited by changing its default setting in the configuration script `$ATMENSETC/AtmEnsConfig.csh`. Bear in mind that not only the members of the ensemble AGCM forecasts will run more slowly , but the archiving mechanism will be overloaded when new output streams are added.

4. *Can I rename the ensemble ADAS script,* `atm_ens.j`*?*
   Yes. Just like the main ADAS script `g5das.j` can be renamed at the user's will, so can the script driving the ensemble. If this is done, both these scripts must be edited and changed accordingly since they contain their own names and each other's name. Renaming of the main ADAS script is usually done during the (fv)setup procedure. However, renaming of the `atm_ens.j` must be done by hand.

5. *Why aren't the observers run in conjunction with the atmospheric AGCM integrations?* GEOS AGCM and the GSI observer are hooked through an ESMF gridded-component so that it is conceivable to invoke the GSI observer while running the AGCM (this has been developed for the GEOS 4D-Var system) – call it the online observer. This question is thus rather pertinent, since the online observer would provide for a more efficient way of running the ensemble of observers in the ensemble ADAS. Unfortunately, details related to how GSI does its quality control based on a certain snapshot of the background fields are such that an offline (regular) observer always ends up taking in more observations than its online counterpart. We may revisit this at some point in the future, but for the time being the most effective way of maximizing usage of the observations is by running offline observers as currently done in the ensemble ADAS (and in our experimental 4D-Var, for that matter).

6. *What happens inside the work directory?* As with the regular ADAS, the ensemble ADAS works within its own reserved space area. Lots of things happen inside this area. Assuming the default options are being exercised, the most important sequence of events follows the flow diagram shown in Fig. 24:

   (a) Links to the NMC-like perturbations are created within a subdirectory named `addperts` created inside the work directory. A subdirectory of this, named `tmperts`, is used while removal of the mean from these perturbations is taking place. Once this process completes, the perturbations used in the additive inflation will reside under `addperts`.

[42] Usually set to `$FVHOME/atmens/RST`.

(b) A directory `ensmean` is created inside the work area and observations are brought from the archive into this directory; links are also created pointing to the mean background files normally sitting under `$FVHOME/atmens/ensmean`; when the observer mean is completed, directories for each member are created inside the work area `$ENSWORK`, and the post-quality-control observation files generated by the mean observer are linked inside each member directory together with the corresponding backgrounds found under `$FVHOME/atmens`.

(c) Once the observers are finished, all observer GSI diagnostics output files and corresponding background files are directly linked inside the main work area directory. The EnKF runs here, and analysis files are originally written out in this directory.

(d) Completion of the ensemble analysis triggers a move of the analyses files from the work area into a subdirectory of this area named `updated_ens`. Inside this directory, each analysis member is placed in its own subdirectory, such as, *mem001, mem002, etc*; links are then created back to the original member directories under the work area. Note that at this stage we have directories named *mem001, mem002, etc*, under the work area, as well as under `updated_ens`, though these are physically distinct directories.

(e) The ensemble mean analysis calculation can now take place, and the resulting mean analysis file is placed under the subdirectory `updated_ens/ensmean`; similary, second-order statistics are placed under `updated_ens/ensrms`.

(f) With availability of the ensemble mean analysis, the member analyses re-centering and inflation can take place inside `$ENSWORK/updated_ens`. Ultimately, the EnKF analyses are overwritten. Notice that due to the links created earlier, the original member directories under `$ENSWORK` see the re-centered and inflated updated analysis files.

(g) Creation of the IAU-forcing terms for each member now takes place inside the member directories under `$ENSWORK`.

(h) Links to the AGCM restart files are created from their original location `$FVHOME/atmens` into the work area member directories inside `$ENSWORK`. Edited resource files and links to boundary condition files are also placed inside each member directory, and the AGCM ensemble is then integrated forward.

(i) The output of each member AGCM integration is moved to the corresponding location under `$ENSWORK/updated_ens`. This directory now has a completely new ensemble with information needed for the next analysis cycle.

(j) At this point, the main script swaps the old ensemble with the new in the original location. That is, as discussed before, this is what happens in the main driver:
/bin/mv $ATMENSLOC/atmens $ATMENSLOC/atmens4arch.${nymdb}_${hhb}
/bin/mv $FVWORK/updated_ens $ATMENSLOC/atmens

(k) Post-processing of the output from the AGCM takes place; mean and other statistics from the members of the ensemble are calculated.

(l) The main job script driving the ensemble ADAS can now launch the hybrid ADAS script, as well as the archiving script that works to permanently store the members from the *previous cycle* (under `$ATMENSLOC/atmens4arch.$nymdb_$hhb`).

7. *What else should I watch out for?*

- When creating an initial ensemble from scratch, and placing it under `$HYBRIDGSI` (i.e., `$FVHOME/run/atmens`), remember to touch a hidden file named `.no_archiving` inside this directory. This is required to prevent the archiving procedure of the central ADAS from looking inside this directory for files to be archived. Remember that each subdirectory of `$HYBRIDGSI`, holding each member of the ensemble, will have files with typical ADAS names,

for example, files fitting a template of the type `%s.bkg.eta.%y4%m2%d2_%h2z.nc4` will be under each member directory. Once the archiving procedure sees these files, it will work to place them in the archive, possibly overwriting whatever the ADAS has placed there. The presence of the hidden file `.no_archiving` in the top directory of a chain of directories is enough to tell the archiving procedure to ignore the directory and its subdirectories.

- Unlike the scripts running the (hybrid) ADAS, the scripts running the *ensemble* ADAS never copy their resource files into the working area. Therefore, if you decide to make changes to any of the resource files under the experiment directory defined by `ATMENSETC` while the ensemble job is running, the changes will be instantly picked up by the run, the exception being changes made to `AtmEnsConfig.csh`. We advise strongly against making such changes while the job is running unless you really understand the potential consequences.

# 12 Future Releases

*Whereof one cannot speak, thereof one must be silent.*

Tractatus Logico-Philosophicus, L. Wittgenstein, 1918.

The worth of a software system lies in its flexibility and friendliness to its users and in its clarity to its developers. This second release of the GEOS Ensemble ADAS and its ability to link up with the regular GEOS ADAS to form the GEOS Hybrid Ensemble-Variational Atmospheric Data Assimilation System has now benefitted from its use in supporting the GMAO Forward Processing System for the past few years. The software, and its two developers, have also benefitted from the many GMAO colleagues who have used it in their research and applications and have helped us correct and improve upon the software's initial release of 2013. This being a live system, new features have been added to the present release, and we expect that continual feedback from users will allow us to fix any new or pending issues that might be found. We recognize there are still weaknesses in the system, and we plan to continue to address them as effectively as possible. The original version of this document listed a number of items for "Future Work". In revising this document we are happy to remove some of those items since they are now available features in the current release. Examples are the addition of the Scheduler, mechanisms to allow running the GEOS Atmospheric Ensemble Forecasting System, and mechanisms to allow running the GEOS Ensemble Forecast Sensitivity and Observation Impact. Plenty of work, however, still remains to be done. An updated list follows below.

**Scripts.** Polyglot programmers will find coding in c-shell to be limited and even annoying. Though we recognize the power of modern languages such as Perl and Python, we feel c-shell provides the clarity that other programming languages lack. However, as new flexibilities and options are added to the scripts, we will be looking into promoting some of them to more modern (perhaps object-oriented) languages.

**Event Log.** Another priority to work on relates to the development of an Event Log mechanism that looks inside the work area and is capable of telling the user what processes are running at any given time of the ensemble ADAS integration. Indeed, such an Event Log is expected to give proper hints into what might have failed in cases when the run stops.

**Archiving.** The archiving mechanism tends to be rather time consuming. In the next couple of months we plan to have another look at the mechanism we presently use. Perhaps joint work with some of our colleagues will lead us into more efficient ways of storing the massive amount of information presently generated by the ensemble, not to mention the potential information presently not put out by the ensemble.

**AGCM Initialization.** One contributor to the large volume of data handled by the archiving mechanism is the large number of restart files (initial conditions) required by the model. We plan to test a version of the ensemble ADAS that essentially bootstraps some of the physics at each cycle. This has the potential to reduce dramatically the number of restarts to be carried along by each member. As we understand, some other centers running hybrid system have a similar strategy. Needless to say, the central (hybrid) ADAS restarts will continue to be handled as a full set.

**Environment Variables to Revisit.** The environment variable `AENSADDINFLOC` controls where the perturbations used in the additive inflation procedure are to be placed. It is presently set in the ensemble configuration file `AtmEnsConfig.csh`, giving the user freedom to change it at will. In reality, this is more of an internal variable that only the ensemble procedure should have control over. Future releases will revisit this.

**LETKF Analysis.** The EnKF software from J. Whitaker used in GEOS EnADAS supports not only the EnSRF option as seen throughout this exposition but also an LETKF option. Since the LETKF does not use serial processing of observations, but rather assimilates observations falling within patches of

the grid, it has the advantage of having its results not be dependent on the order in which observations are assimilated. For this reason, it would be desirable to experiment with the LETKF. Although it can be easily triggered with a simple setting in its namelists, in practice, since the LETKF is more memory demanding than the EnSRF, tuning of resources is required to robustly enable this option in GEOS EnADAS.

**Ensemble Aerosol Analysis.** In Secs. 2 and 4.1.2 we briefly mentioned how, in GEOS EnADAS, the ensemble of model integrations presently relies on the central, PSAS-based, analysis of aerosol optical depth (AOD) to feed into the Local Displacement Ensemble updates of aerosol concentrations taking place in each ensemble member. Work has already been done (Buchard et al. pers. comm.) to enable the EnKF/EnSRF to analyze either AOD or aerosol concentrations. Upcoming releases of GEOS Hybrid EVADAS are expected to incorporate this EnKF-based AOD analysis so each member ensemble will have its own instance of analyzed AOD, thus contributing to improved spread in the aerosol concentrations generated by the ensemble. Ultimately, the plan is to have the PSAS analysis replaced with an ensemble-based analysis of aerosols even when it relates to the deterministic system.

**EnKF cycled satellite bias estimates.** As mentioned in Sec. 8.5, the EnKF software of J. S. Whitaker is capable of estimating satellite biases. However, our scripts are not yet enabled to cycle these estimates properly. A knob will be added to allow this feature and permit fully independent experimentation with ensemble-only strategies. This is obviously only a concern when the data assimilation is run in pure ensemble mode, without any connections with the hybrid cycle.

**Augmenting and updating the database of NMC-like perturbations.** The current database of 48-minus-24-hour perturbations relied on by the additive inflation mechanism, as well as by the NMC-method estimation procedure for deriving the climatological background error covariance needed by the hybrid GSI, covers a single one-year period of forecasts derived from the FP system version 5.7.2. This is a version not too different from that which supported MERRA-2. However, considering that (i) the present version of the system is considerably different from 5.7.2, and (ii) there will be a need to support a reanalysis system capable of using the Hybrid GEOS system, it will be necessary to reconstruct the database with more recent forecasts from GEOS FP version 5.17.0, and to expand the database to cover most of the 30-plus-year reanalysis period. The latter might be simplified somewhat by having the database cover only the periods for which there is a substantial observing system change within the 30-plus years period (such as before and after introduction of SSMI observations, before and after introduction of MSU observations, and other such periods, with special care given to overlapping periods).

**Facilitating OSE experimentation.** As briefly mentioned in the introduction, among the options of the ensemble analysis within the GEOS EnADAS (see Figs. 1 and 2) is the capability of running an Ensemble of Data Assimilation Systems. Though this feature is not yet fully capable of supporting a cycled version of something like an ensemble of Hybrid 4D-EnVar's, the hybrid analysis can run in replayed mode (when the ensemble comes from an existing experiment) and thus it can be used to support Observation System Experiments. Enabling an OSE framework is rather trivial using the EDA capability. This feature exercises the full machinery of the GEOS EnADAS, without coupling the "ensemble" of OSE's with a central deterministic system. Indeed, in this form, the deterministic run can be made to function as the control experiment, where all observations are assimilated, and the individual members of the ensemble (which are never coupled in any way) can be set to handle a separate selection of instruments and observations forming typical partitioning in OSEs. For example, member one can be set to assimilate only conventional observations, member two can be set to assimilate only AMSUA, and so on. In such cases, the deterministic and ensemble system should run at the same resolution for consistency. In a sense, everything needed for running OSEs within the EDA capability is in place, but this is placed here as Future Work, just because no actual experiments have yet been done.

**Enabling EDA.** The capability of running a fully cycled EDA is not far into the future. Setting this up in GEOS EnADAS requires minor additions in a place or two. Specifically, when the EDA entails an ensemble of hybrid analyses, it is necessary to define how the members of the ensemble are used in each Hybrid member making sure no redundancy is built into the Hybrid variational analyses of the members. This is another feature rather simple to implement as a mild expansion of existing features. Having EDA as an easily set up option within GEOS Hybrid EVADAS is planned for upcoming releases.

This page intentionally left blank.

# 13 Acknowledgments

We start by thanking Jeffrey Whitaker of ESRL/NOAA for providing us with the EnKF software that is core to the EnADAS development and for being a very helpful consultant in the early stages of its development. We also thank David Parrish, from NCEP/NOAA, for his original implementation of a hybrid capability in GSI, and Daryl Kleist and Rahul Mahajan for helping us compare GMAO and NCEP hybrid implementations. Much appreciation and many thanks go to the GMAO users of GEOS ADAS for their help in finding unfriendly features in our constructs while trying to experiment within GEOS Hybrid EVADAS. We thank them in particular for their patience in waiting for fixes and changes to come their way. Thanks are also due to the GMAO Monitoring Group for their diligence in helping evaluate the multiple intermediate versions of this system produced as it was developed and for spotting, at times, puzzling behavior that needed our close attention. Special thanks are due to Mark Solomon and Robert Lucchesi for their willingness and dedication in learning the intricacies and new features behind the underlying EnADAS supporting the Hybrid ADAS. We thank Randal Koster for revising this document and providing comments that help improve its readability. Many thanks are also due to the management of the NASA Center for Climate Similation (NCCS) for making sure computing resources were made available to us to allow for the development and implementation of the upgrades here into our Forward Processing (FP) System. Finally, we thank the GMAO management for their support throughout the development of this system. Our development started under Michele Rienecker, and gladly before she ended her term as Head of GMAO, we delivered the first version of the Hybrid 3D-Var to the GMAO Quasi-Operational Group and upgraded the FP system from 3D-Var to Hybrid 3D-Var. The January 2017 upgrade of our FP system to Hybrid 4D-EnVar happened under Steven Pawson. Both Michele and Steven have provided us encouragment for this development and have "tortured" us enough to write and update this document.

This page intentionally left blank.

# References


Ancell B, Hakim GJ. 2007. Comparing adjoint- and ensemble-sensitivity analysis with applications to observation targeting. *Mon. Wea. Rev.* **135**: 4117–4134.

Anderson JL. 2009. Spatially and temporally varying adaptive covariance inflation for ensemble filters. *Q. J. Royal Meteorol. Soc.* **137**: 72–83.

Bannister RN. 2008. A review of forecast error covariance statistics in atmospheric variational data assimilation. I: Characteristics and measurements of forecast error covariances. *Q J Royal Meterol Soc* **134(B)**: 1951–1970.

Bannister RN. 2017. A review of operational methods of variational and ensemble-variational data assimilation. *Q J Royal Meterol Soc* **143(B)**: 607–633.

Bishop CH, Hodyss D. 2011. Adaptive ensemble covariance localization in ensemble 4D-VAR state estimation. *Q. J. Royal Meteorol. Soc.* **139**: 1241–1255.

Bloom SC, Takacs LL, da Silva AM, Ledvina D. 1996. Data assimilation using incremental analysis updates. *Mon. Wea. Rev.* **124**: 1256–1271.

Bowler NE, Clayton AM, Jardak M, Lee E, Lorenc AC, Piccolo C, Pring SR, Wlasak MA, Barker DM, Inverarity GW, Swinbank R. 2017. Inflation and localization tests in the development of an ensemble of 4D-ensemble variational assimilations. *Q. J. Royal Meteorol. Soc.* **143A**: 1280–1308.

Buehner M, Houtekamer PL, Charette C, Mitchell HL, He B. 2010a. Intercomparison of variational data assimilation and the ensemble Kalman filter for global deterministic NWP. Part I: Description and single-observation experiments. *Mon. Wea. Rev.* **138**: 1550–1566.

Buehner M, Houtekamer PL, Charette C, Mitchell HL, He B. 2010b. Intercomparison of variational data assimilation and the ensemble Kalman filter for global deterministic NWP. Part II: One-month experiments with real observations. *Mon. Wea. Rev.* **138**: 1567–1586.

Charron M, Pellerin G, Spacek L, Houtekamer PL, N Gagnon HLM, Michelin L. 2010. Toward random sampling of model error in the Canadian Ensemble Prediction System. *Mon. Wea. Rev.* **138**: 1877–1901.

Chou MD, Suarez MJ. 1999. A solar radiation parameterization for atmospheric studies. NASA Tech. Memo. 104606-15, NASA Goddard Space Flight Center.

Clayton AM, Lorenc AC, Barker DM. 2013. Operational implementation of a hybrid ensemble/4D-Var global data assimilation system at the Met Office. *Q. J. Royal. Meteorol. Soc.* **139B**: 1445–1461, doi: 10.1002/qj.2054.

Clough S, Shephard M, Mlawer E, Delamere J, Iacono M, Cady-Pereira K, Boukabara S, PDBrown. 2005. Atmospheric radiative transfer modeling: a summary of the AER codes. *Journal of Quantitative Spectroscopy & Radiative Transfer* **91**: 233–244.

Cohn SE, da Silva A, Guo J, Sienkiewicz M, Lamich D. 1998. Assessing the effects of data selection with the DAO physical-space statistical analysis system. *Mon. Wea. Rev.* **126**: 2913–2926.

Colarco P, da Silva A, Chin M, Diehl T. 2010. Online simulations of global aerosol distributions in the NASA GEOS-4 model and comparisons to satellite and ground-based aerosol optical depth. *J. Geophys. Res.* **115**: D14 207, doi:doi:10.1029/2009JD012820.

Collins N, Theurich G, Deluca C, Suarez M, Trayanov A, Balaji V, Li P, W Yang CH, da Silva A. 2005. Design and implementation of components in the Earth System Modeling Framework. *Intl. J. High Perform. Comput. Appl.* **19(3)**: 341–350.

Courtier P. 1997. Dual formulation of four-dimensional variational assimilation. *Q. J. Royal Meteorol. Soc.* **123**: 2449–2461.

Daescu DN, Todling R. 2009. Adjoint estimation of the variation in model functional output due to the assimilation of data. *Mon. Wea. Rev.* **137**: 1705–1716.

Daescu DN, Todling R. 2010. Adjoint sensitivity of the model forecast to data assimilation system error covariance parameters. *Q. J. Royal Meteorol. Soc.* **136**: 2000–2012.

Derber JC, Rosati A. 1989. A global oceanic data assimilation technique. *J. Phys. Oceanogr.* **19**: 1333–1347.

Derber JC, Wu WS. 1998. The use of TOVS could-cleared radiances in the NCEP SSI analysis system. *Mon. Wea. Rev.* **126**: 2287–2299.

Desroziers G, Berre L, Chapnik B, Poli P. 2005. Diagnosis of observation, background and analysis-error statistics in observation space. *Q. J. Royal Meteorol. Soc.* **131**: 3385–3396.

El Akkraoui A, Trémolet Y, Todling R. 2013. Preconditioning of variational data assimilation and the use of a bi-conjugate gradient method. *Q. J. Royal. Meteorol. Soc.* **139**: 731–741.

Errico RM, Gelaro R, Novakovskaia E, Todling R. 2007. General characteristics of stratospheric singular vectors. *Meteorologische Zeitschrift* **16**: 621–634.

Eyre JR. 2016. Observation bias correction schemes in data assimilation systems: a theoretical study of some of their properties. *Q J Royal Meteorol Soc* **142**: 2284–2291.

Fisher M. 1998. Minimization algorithms for variational data assimilation. Technical Report Seminary on Recent Developments in Numerical Methods for Atmospheric Modelling, ECMWF.

Fletcher S. 2017. *Data Assimilation for the Geosciences*. Elsevier: New York City, first edn.

Fortin V, Baza MA, Anctil F, Turcotte R. 2014. Why should ensemble spread match the RMSE of the ensemble mean? *J. Hydrometeorology* **15**: 1708–1713.

Gauthier P, Thépaut JN. 2001. Impact of the digital filter as a weak constraint in the preoperational 4DVAR assimilation system of météo france. *Mon. Wea. Rev.* **129**: 2089–2102.

Gelaro R, coauthors. 2017. The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *J. Climate* **30**: 5419–5454.

Gelaro R, Langland RH, Pellerin S, Todling R. 2010. The THORPEX observation impact inter-comparison experiment. *Mon. Wea. Rev.* **138**: 4009–4025.

Giering R, Kaminski T, Todling R, Errico R, Gelaro R, Winslow N. 2017. Generating tangent linear and adjoint versions of NASA/GMAO's Fortran-90 global weather forecast model. *In H. M. Bücker, G. Corliss, P. Hovland, U. Naumann, and B. Norris, editors, Automatic Differentiation: Applications, Theory, and Implementations, Lecture Notes in Computational Science and Engineering* **50**: 275–284.

Groff D. 2017. Assessment of ensemble forecast sensitivity to observation (EFSO) quantities for satellite radiances assimilated in the 4DEnVar GFS. *Joint Center for Satellite Data Assimilation Quarterly* **54**: 1–5.

Gürol S, Weaver AT, Moore AM, Piacentini A, Arango AG, Gratton S. 2014. B-preconditioned minimization algorithms for variational data assimilation with the dual formulation. *Q. J. Royal Meteorol. Soc.* **140B**: 539–556.

Ham YG, Rienecker MM, Suarez MJ, Vikhliaev Y, Zhao B, Marshak J, Vernieres G, Schubert SD. 2013. Decadal prediction skill in the GEOS-5 forecast system. *Climate Dynamics* **42**: 1–20.

Hamill TM, Snyder C. 2000. A hybrid ensemble Kalman filter-3D variational analysis scheme. *Mon. Wea. Rev.* **128**: 2905–2919.

Hamill TM, Whitaker JS. 2005. Accounting for the error due to unresolved scales in ensemble data assimilation: A comparison of different approaches. *Mon. Wea. Rev.* **133**: 3132–3147.

Holdaway D, Errico R, Gelaro R, Kim J. 2014. Inclusion of linearized moist physics in NASA's Goddard Earth Observing System Data Assimilation tools. *Mon. Wea. Rev.* **142**: 414–433.

Hollingsworth A, Lönnberg P. 1989. The verification of objective analyses: Diagnostics of analysis system performance. *Meteorol. Atmos. Phys.* **40**: 3–27.

Hu M, Zhou C, Shao H, Stark D, Newman K. 2016. Advanced GSI User's Guide. Tech. memo., Developmental Testbed Center, UCAR, Boulder, URL https://dtcenter.org/com-GSI/users/docs/users_guide/AdvancedGSIUserGuide_v3.5.0.0.pdf.

Isaksen L, Bonavita M, Buizza R, Fisher M, Haseler J, Leutbecher M, Raynaud L. 2010. Ensemble of Data Assimilations at ECMWF. Technical Report Tech. Memo. 636, ECMWF.

Kailath T. 1968. An innovations control approach to least square estimation — Part I: Linear filtering in additive white noise. *IEEE Trans. Automat. Control* **AC-13**: 646–655.

Kleespies TJ, van Delst P, McMillin LM, Derber JC. 2004. Atmospheric transmittance of an absorbing gas. 6. OPTRAN status report and introduction to the NESDIS/NCEP Community Radiative Transfer Model. *Appl. Opt.* **43**: 3103–3109.

Kleist DT. 2011. Assimilation of tropical cyclone advisory minimum sea level pressure in the NCEP Global Data Assimilation System. *Wea. Forecasting* **26**: 1085–1091.

Kleist DT. 2012. An evaluation of hybrid variational-ensemble data assimilation for the NCEP GFS. Ph. d. thesis, University of Maryland, URL http://www.emc.ncep.noaa.gov/gmb/wd20dk/docs/phd/DarylKleist_PhDThesis_Revised.pdf.

Kleist DT, Mahajan R, Desroziers G, Berre L, Buehner M, Lorenc A, Isaksen L, Trémolet Y, Bonavita M, Potthast R, Kadowaki T. 2018. Survey of data assimilation implementations for global numerical weather prediction at operational meteorological centers. *Q J R Meteorol Soc* : In preparation.

Kleist DT, Parrish DF, Derber JC, Treadon R, Errico RM, Yang R. 2009a. Improving incremental balance in the GSI 3DVAR analysis system. *Mon. Wea. Rev.* **137**: 1046–1060.

Kleist DT, Parrish DF, Derber JC, Treadon R, Wu WS, Lord S. 2009b. Introduction of the GSI into the NCEP Global Data Assimilation System. *Wea. Forecasting* **24**: 1691–1705, doi:http://dx.doi.org/10.1175/2009WAF2222201.1.

Koster RD, Suarez MJ, Ducharne A, Stieglitz M, Kumar P. 2000. A catchment-based approach to modeling land surface processes in a GCM, Part I: model structure. *J. Geophys. Res.* **105(D20)**: 24 809–24 822.

Langland RH, Baker NL. 2004. Estimation of observation impact using the NRL atmospheric variational data assimilation adjoint system. *Tellus* **56A**: 189–201.

Lewis JM, Lakshmivarahan S, Dhall S. 2006. *Dynamic data assimilation: A least squares approachr*. Cambridge University Press: Cambridge, UK.

Li H, Liu J, Kalnay E. 2010. Correction of 'estimating observation impact without adjoint model in an ensemble Kalman filter'. *Q. J. Royal Meteorol. Soc.* **136**: 1652–1654.

Lin SJ. 2004. A vertically Lagrangian finite-volume dynamical core for general circulation models. *Mon. Wea. Rev.* **132**: 2293–2307.

Liu J, Kalnay E. 2008. Estimating observation impact without adjoint model in an ensemble Kalman filter. *Q. J. Royal Meteorol. Soc.* **134**: 1327–1335.

Lorenc AC. 2003. The potential of the ensemble Kalman filter for NWP – A comparison with 4D-Var. *Q. J. R. Meteorol. Soc.* **129**: 3183–3203.

Lorenc AC, Bowler NE, Clayton AM, Pring SR. 2015. Comparison of Hybrid-4DEnVar and Hybrid-4DVar data assimilation methods for global NWP. *Mon. Wea. Rev.* **143**: 212–229.

Lupu C, Gauthier P, Laroche S. 2011. Evaluation of the impact of observations on analyses in 3D- and 4D-Var based on information content. *Mon. Wea. Rev.* **139**: 726–737.

Marchok TP. 2010. How the NCEP tropical cyclone tracker works. *Mon. Wea. Rev.* **138**: 4509–4522.

Massart SM, Pajot B, Piacentini A, Pannekoucke O. 2010. On the merits of using a 3D-FGAT assimilation scheme with an outer loop for atmospheric situations governed by transport. *Preprints, 25th Conf. on Hurricanes and Tropical Meteorology, San Diego, CA, Amer. Meteor. Soc.* : 21–22.

McCarty W, Coy L, Gelaro R, Huang A, Merkova D, E B Smith MS, Wargan K. 2016. A solar radiation parameterization for atmospheric studies. NASA Tech. Memo. 104606-46, NASA Goddard Space Flight Center.

Miyoshi T. 2011. The Gaussian approach to adaptive covariance inflation and its implementation with the local ensemble transform Kalman filter. *Mon. Wea. Rev.* **139**: 1519–1535.

Moorthi S, Suarez MJ. 1992. A parameterization of moist convection for general-circulation models. *Mon. Wea. Rev.* **120**: 978–1002.

Ota Y, Derber JC, Kalnay E, Miyoshi T. 2013. Ensemble-based observation impact estimates using NCEP GFS. *Tellus* **65**: 1–14.

Palmer T, Buizza R, Doblas-Reyes F, Jung T, Leutbecher M, Shutts G, Steinheimer M, Weisheimer A. 2009. Stochastic parameterization and model uncertainty. Technical Report Tech. Memo. 598, ECMWF.

Parrish DF, Derber JC. 1992. The National Meteorological Center's spectral statistical interpolation analysis system. *Mon. Wea. Rev.* **120**: 1747–1763.

Polavarapu S, Ren S, Clayton A, Sankey D, Rochon Y. 2004. On the relationship between Incremental Analysis Updating and Incremental Digital Filtering. *Mon. Wea. Rev.* **132**: 2495–2502.

Putman WM, Lin SJ. 2007. Finite-volume transport on various cubed-sphere grids. *J. Comput. Phys.* **227**: 55– 78, doi:10.1016/j.jcp.2007.07.022.

Rabier F, Järvinen H, Klinker E, Mahfouf JF, Simons A. 2000. The ECMWF operational implementation of four-dimensional variational assimilation. I: Experimental results with simplified physics. *Q. J. Royal Meteorol. Soc.* **126A**: 1143–1170.

Randles RC, da Silva AM, Buchard V, Colarco PR, Darmenov A, Govindaraju R, Smirnov A, Holben B, Ferrare R, Hair J, Shinozuka Y, Flynn CJ. 2017. An improved in situ and satellite SST analysis for climate. *J. Climate* **30**: 6824–6850.

Rawlins F, Ballard S, Bovis K, Clayton A, Li D, Inverarity G, Lorenc A, Payne T. 2007. The Met Office global four-dimensional variational data assimilation scheme. *Q. J. Royal Meteorol. Soc.* **133**: 347–362.

Rienecker MM, coauthors. 2008. The GEOS-5 Data Assimilation System. Documentation of versions 5.0.1 and 5.1.0, and 5.2.0. Technical Report Series on Global Modeling and Data Assimilation NASA/TM-2008-104606/Vol 27, NASA Goddard Space Flight Center.

Rienecker MM, coauthors. 2011. MERRA: NASAs Modern-Era Retrospective Analysis for Research and Applications. *J. Climate* **24**: 3624–3648.

Stieglitz M, Ducharne A, Koster RD, Suarez MJ. 2001. The impact of detailed snow physics on the simulation of snow cover and subsurface thermodynamics at continental scales. *J. Hydrometeor* **2**: 228–242.

Takacs LL, Suárez MJ, Todling R. 2016. Maintaining atmospheric mass and water balance in reanalyses. *Q. J. Royal Meteorol. Soc.* **142B**: 1565–1573.

Takacs LL, Suárez MJ, Todling R. 2018. The stability of Incremental Analysis Update for global models. *Mon. Wea. Rev.* : In preparation.

Todling R. 2013. Comparing two approaches for assessing observation impact. *Mon Weather Rev* **141**: 1484–1505, doi:http://dx.doi.org/10.1175/MWR-D-12-00100.1.

Todling R, El Akkraoui A. 2013. The GMAO Hybrid Ensemble-Variational Atmospheric Data Assimilation System: Version 1.0. Gmao office note, NASA, URL https://gmao.gsfc.nasa.gov/intranet/personnel/rtodling/dasdev/AtmosEnsADAS_Rel1.pdf.

Trahan S, Sparling L. 2012. An analysis of NCEP tropical cyclone vitals and potential effects on forecasting models. *Wea. Forecasting* **27**: 744–756.

Trémolet Y. 2007a. First-order and higher-order approximations of observation impact. *Meteorologische Zeitschrift* **16**: 693–694.

Trémolet Y. 2007b. Incremental 4D-Var convergence study. *Tellus* **59**: 706–718.

Trémolet Y. 2008. Computation of observation sensitivity and observation impact in incremental variational data assimilation. *Tellus* **60A**: 964–978.

Wang X, Snyder C, Hamill TM. 2007. On the theoretical equivalence of differently proposed ensemble/3D-Var hybrid analysis schemes. *Mon. Wea. Rev.* **135**: 222–227.

Whitaker JS, Hamill TM, Wei X, Song Y, Toth Z. 2008. Ensemble data assimilation with the NCEP Global Forecast System. *Mon. Wea. Rev.* **136**: 463–482.

Woollen JS, Kalnay E, Gandin L, Collins W, Saba S, Kistler R, Kanamitsu M, Chelliah M. 1994. Quality control in the reanalysis system. 10th Conf. on Numerical Weather Prediction, Portland, OR. Amer. Meteor. Soc., pp. 13–14.

Wu W, Purser RJ, Parrish DF. 2002. Three dimensional variational analysis with spatially inhomogeneous covariances. *Mon. Wea. Rev.* **130**: 2905–2916.

Zhu Y, Derber J, Collard A, Dee D, Treadon R, Gayno G, Jung JA. 2013. Enhanced radiance bias correction in the National Centers for Environmental Prediction's Gridpoint Statistical Interpolation data assimilation system. *Q. J. Royal Meteorol. Soc.* **140**: 1479–1492.

This page intentionally left blank.

# Appendix A

## Usage command-line of driving scripts behind GEOS EnADAS

This appendix provides the usage command-line of the main scripts controlling GEOS EnADAS. The following is a list of the routines included here:

**acquire_atmens.csh** - acquire pre-existing Atmos-Ensemble

**acquire_ensperts.csh** - acquires NMC-like ensemble perturbations

**atmens_recenter.csh** - recentering script uses files retrieved here

**atmens_addinflation.csh** - apply additive inflation to members of ensemble

**atmens_arch.csh** - prepare to archive atmos-ensemble

**atmens_berror.csh** - create parameterized B-error from ensemble members

**atmens_calcaod.csh** - calculate AOD from Concentration files

**atmens_obsimp0hr.csh** - calculate observation impact on analysis

**atmens_prepgeps.csh** - prepare environment for GEOS ensemble-prediction system (GEPS)

**atmens_prepobsens.csh** - prepare environment for ensemble-FSO calculations

**atmens_restarts.csh** - entry point to ensemble AOD analysis

**atmens_seasonal_dates.csh** - generate random dates falling within season

**atmens_stats.csh** - calculates statistics from ensemble members

**atmens_vtrack.csh** - calculate TC track for each member of ensemble

**atmos_eana.csh** - call desired Ensemble Analysis strategy.

**atmos_eaod.csh** - entry point to ensemble AOD analysis

**atmos_eezy.csh** - driver for simplified scheme

**atmos_egsi.csh** - driver for ensemble of GSI analysis

**atmos_enkf.csh** - driver for EnKF analysis

**atmos_ens2gcm.csh** - construct IAU increment for each member

**edas_scheduler.csh** - scheduler for running EnADAS in parallel

**gcm_ensemble.csh** - run multiple copies of (atmospheric) GCM

**gen_ensbkg.csh** - artificially generates ensemble backgrounds

**gen_ensemble.csh** - artificially generates ensemble

**gen_ensrst.csh** - artificially generates ensemble of GCM restarts

**gen_nmcperts.csh** - generate NMC-perturbations from (OPS) forecasts

**jobmonitor.csh** - monitor events of a parallel section

**makeiau.csh** - create IAU increment for given member

**obsvr_ensemble.csh** - run observer over mean background and each ensemble member

**setobsvr.csh** - prepare observations to be used by observer (based on fvssi)

**obsvr_ensfinal.csh** - run observer over mean analysis to get OmAs

**post_eana.csh** - post ensemble-analysis calculations

**obsvr_ensfinal.csh** - calculate OmAs from mean analysis

**post_efso.csh** - post ensemble-FSO calculations

**post_egcm.csh** - post processing after ensemble of GCMs

**recenter.csh** - recenter ensemble of analyses

**setperts.csh** - set perturbations for additive inflation

**update_ens.csh** - saves updated ensemble member as it completes running

```
NAME

 acquire_atmens.csh  - acquire pre-existing Atmos-Ensemble

SYNOPSIS

 acquire_atmens.csh  expid nymd nhms rcfile

 where
  expid  -  usual experiment name, e.g., b541iau
  nymd   -  analysis date, as in YYYYMMDD
  nhms   -  analysis time, as HHMMSS
  rcfile -  full path name of acquire rc file

DESCRIPTION

  Acquire pre-existing Atmos-Ensemble and make it available
  to ongoing experiment. This is to allow running a hybrid-
  GSI experiment without the need for regenerating the ensemble.

  This script is also used to retrieve the pre-existing ensemble
  when running observation impact.

  Example of valid command line:
    acquire_atmens.csh b541iau 20091018 00 atmens_replay.acq

REQUIRED ENVIRONMENT VARIABLES:

   FVHOME        - location of experiment
   FVROOT        - location of DAS build
   FVWORK        - work directory where ensemble will fall
   GID           - group ID to run job under
   TIMEINC       - analysis frequency (minutes)
   VAROFFSET     - off-time of forecast wrt to 1st synoptic time

RESOURCE FILES

   atmens_replay.acq  - use this in ensemble replay mode
   atmens_asens.acq   - use this when running hybrid adjoint GSI

OPTIONAL EVIRONMENT VARIABLES

   ATMENSLOC     - place where to put acquired ensemble
                   (default: FVWORK)
SEE ALSO
   analyzer  - driver for central ADAS analysis

AUTHOR
  Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    Last modified: 08Apr2013      by: R. Todling
```

```
NAME

 acquire_ensperts.csh  - acquires NMC-like ensemble perturbations

SYNOPSIS

 acquire_ensperts.csh  expid member nymd nhms INCLOC

  where
  expid  -  usual experiment name, e.g., b541iau
  member -  number of member to operate on
  nymd   -  date of analysis, as in YYYYMMDD
  nhms   -  time of analysis, as in HHMMSS
  INFLOC -  location to place perturbations at

DESCRIPTION

  Acquire NMC-like perturbations from database.

REQUIRED ENVIRONMENT VARIABLES

   ATMENSETC - location of ensemble resource files
   ATMENSLOC - location of ensemble, usually FVHOME
   FVHOME    - location of experiment
   FVROOT    - location of DAS build
   FVWORK    - location of DAS work directory

OPTIONAL ENVIRONMENT VARIABLES

   NCSUFFIX  - suffix of hdf/netcdf files (default: nc4)

SEE ALSO
   atmens_recenter.csh - recentering script uses files retrieved here

AUTHOR
  Amal El Akkraoui (Amal.ElAkkraoui@nasa.gov), NASA/GMAO
    Last modified: 08Apr2013      by: R. Todling
```

```
NAME

atmens_addinflation.csh  - apply additive inflation to members of ensemble

SYNOPSIS

atmens_addinflation.csh nmem ftype ensloc

where
  nmem   -  number of members to be created
  ftype  -  file type (e.g., bkg.eta, bkg.sfc, ana.eta)
  ensloc -  location to place generated ensemble

DESCRIPTION

 Apply additive inflation to member analysis.
 NOTE: CURRENTLY NOT USED (additive inflation done via recentering)

Example of valid command line:
atmens_addinflation.csh 10 /archive/u/rtodling/u000_c72/atmens

REQUIRED ENVIRONMENT VARIABLES

   FVROOT    - location of DAS build
   FVHOME    - location of experiment
   EXPID     - experiment name
   ASYNBKG   - frequency of background (minutes)

OPTIONAL ENVIRONMENT VARIABLES

  NCSUFFIX   - SDF variable (default: nc4)

SEE ALSO
 atmens_recenter.csh  -  invoke actual recentering and apply inflation

AUTHOR
  Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    Last modified: 08Apr2013      by: R. Todling
```

```
NAME

  atmens_arch.csh  - prepare to archive atmos-ensemble

SYNOPSIS

 atmens_arch.csh  expid nymd nhms rc

where
  expid  -  usual experiment name, e.g., b541iau
  nymd   -  initial date of ensemble forecast, as in YYYYMMDD
  nhms   -  initial time of ensemble forecast, as HHMMSS
  rc     -  full path name of storage archive file

DESCRIPTION

 Collect files to arquive after Ensemble DAS completes a cycle

 Example of valid command line:
  atmens_arch.csh b541iau 20091018 000000 SOMEDIR/atmens_storage.arc

REQUIRED ENVIRONMENT VARIABLES:

   ARCHLOC       - location of archive, e.g., /archive/u/rtodling
   ATMENSETC     - location of EnKF resource files
   ATMENSLOC     - location of current ensemble
   FVHOME        - location of experiment
   FVROOT        - location of DAS build
   GID           - group ID to run job under
   VAROFFSET     - analysis offset from initial time

OPTIONAL ENVIRONMENT VARIABLES:

   ENSARCH_ALLBKG    - when set, arch all bkg files regardless of hour
                           (Default: arch only central bkg files)
   ENSARCH_FIELDS    - components (list separate by comma), e.g.,
                           eana,ebkg,edia,eprg,erst,eoi0,stat,xtra
                           (Default: stat)
   ENSARCH_WALLCLOCK - location of archive, e.g., /archive/u/rtodling
   NCSUFFIX          - SDF suffix (detault: nc4)

AUTHOR
  Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    Last modified: 08Apr2013      by: R. Todling
```

```
NAME

 atmens_berror.csh  - create parameterized B-error from ensemble members

SYNOPSIS

 atmens_berror.csh  expid nymd nhms idir odir

where
  expid  -  usual experiment name, e.g., b541iau
  nymd   -  date of analysis, as in YYYYMMDD
  nhms   -  time of analysis, as in HHMMSS
  idir   -  location of ensemble
  odir   -  location where to link output B-error file to
              (named: berror_stats)

DESCRIPTION

  This is a driver for running calcstats over the members of the
atmospheric ensemble

Example of valid command line:
atmens_berror.csh b541iau 20091019 000000 somedir someotherdir

REQUIRED RESOURCE FILES

OPTIONAL RESOURCE FILES
   atmens_berror.rc - when found in idir allow for B-error to be
                      created at desired resolution

REQUIRED ENVIRONMENT VARIABLES

   FVHOME          - location of experiment
   FVROOT          - location of DAS build
   MPIRUN_CALCSTATS - controls executable for B-err generation

OPTIONAL ENVIRONMENT VARIABLES

   BERROR_FROMENS - when specified, replaces berror_stats file
                    with that produced here

REMARKS

SEE ALSO
  calcstats.x - program doing actual work
  ut\_atmens\_berror.j - off-line unit tester

AUTHOR
  Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    Last modified: 22Dec2016      by: R. Todling
```

```
NAME

 atmens_calcaod.csh  - calculate AOD from Concentration files

SYNOPSIS

 atmens_calcaod.csh  expid nymd nhms conc aod workdir

where
  expid  -  usual experiment name, e.g., b541iau
  nymd   -  date of analysis, as in YYYYMMDD
  nhms   -  time of analysis, as in HHMMSS
  conc   -  Concentration file label (e.g., abkg)
  aod    -  AOD file label (e.g., aod_f)
  workdir-  frequency of AOD analysis, as in HHMMSS

DESCRIPTION

   Uses the background or analyzed aerosol concentration files
   to produce calculate total aerosol optional depth (AOD).
   Though in principle the offline calculate performed here
   should reproduce the online calculation done within the
   atmospheric model, in practice differing resolution betwen
   the underlying model and the aerosol file provided to this
   procedure will lead to small differences (e.g.,
   cubed AGCM, lat/lon history).

 Example of valid command line:
 atmens_calcaod.csh b541iau 20091019 000000 030000 aana aod_a \
                   /wrkdirname

REQUIRED ENVIRONMENT VARIABLES

 FVHOME        - location of experiment
 FVROOT        - location of DAS build

OPTIONAL ENVIRONMENT VARIABLES

 NCSUFFIX          - suffix of hdf/netcdf files (default: nc4)

RESOURCE FILES
  Chem_MieRegistry.rc, GAAS_Mie.rc

SEE ALSO
  atmos_eaod.csh

AUTHOR
  Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    Last modified: 26Mar2017      by: R. Todling
```

```
NAME

 atmens_obsimp0hr.csh  - calculate obs impact on analysis

SYNOPSIS

 atmens_obsimp0hr.csh  expid nymd nhms anadir

where
  expid     -  usual experiment name, e.g., b541iau
  nymd      -  date of analysis, as in YYYYMMDD
  nhms      -  time of analysis, as in HHMMSS

DESCRIPTION
   This procedures calculates observation impacts on the
 analysis (i.e, impact on the 0-hour forecast). First the
 OMBs are converted from diag to ODS; then the OMAs are
 are converted to ODS; and finally the impacts are produced
 and saved in a set of ODS files that are then available for
 archiving.
   The presence of the file odsmatch.rc under ATMENSETC
 serves as a trigger of this procedure.

 Example of valid command line:
 atmens_obsimp0hr.csh b541iau 20091019 000000

REQUIRED RESOURCE FILES

  odsmatch.rc  - required by odsselect (diag2ods)

REQUIRED ENVIRONMENT VARIABLES

   ATMENSETC       - location of resource files
   FVHOME          - location of experiment
   FVROOT          - location of DAS build
   FVWORK          - location of work directory

OPTIONAL ENVIRONMENT VARIABLES

   NCSUFFIX         - suffix of hdf/netcdf files (default: nc4)
   ENSPARALLEL      - when set, runs all ensemble components in
                      parallel (default: off)
   AENS_OBSVR_DSTJOB- distribute multiple works within smaller
                      jobs
   OBSVR_WALLCLOCK  - wall clock time to run observer
                        (default 1:00:00)
   OBSVR_QNAME      - name of queue
                        (default: NULL, that is, let pbs pick)
```

```
SEE ALSO

  diag2ods  - converts GSI diag files to ODS

AUTHOR
  Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    Last modified: 15Apr2013      by: R. Todling
```

```
NAME

 atmens_prepgeps.csh  - prepare environment for GEOS
                        ensemble-prediction system (GEPS)

SYNOPSIS

 atmens_prepgeps.csh  expid nymd nhms atype action

where
  expid  -  usual experiment name, e.g., b541iau
  nymd   -  date of analysis, as in YYYYMMDD
  nhms   -  time of analysis, as in HHMMSS
  atype  -  analysis type (ana or niana)
  action -  the following are valid:
            setrc  - to generate acquire file with all needed for EFSO
            null   - to actually create all links and resolution conversions

DESCRIPTION

   This procedure is responsible for converting the output of the
 EnKF backward integration into ODS and producing whetever other
 diagnostic and statistic desired from the EFSO procedure.

 Example of valid command line:
 atmens_prepgeps.csh b541iau 20091019 000000 ana setrc

REQUIRED ENVIRONMENT VARIABLES

   AENSTAT_MPIRUN   - mp_stats MPI command line
   ATMENSETC        - location of experiment
   FVWORK           - location of work directory

OPTIONAL ENVIRONMENT VARIABLES

   ATMENS_GEPS_RECENTER 1: use to recenter ensemble analysis
                        (default: 0)
   NCSUFFIX          - suffix of hdf/netcdf files (default: nc4)
   DATADIR           - location where original data resides
                       (default: /archive/u/user)
   SRCEXPID          - original experiment (use when other than expid)
                       (default: expid)

SEE ALSO
  ut_prepobsens.csh - unit tester for this procedure

AUTHOR
  Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    Created modified: 01Apr2017   by: R. Todling
    Last modified: 16Apr2017      by: R. Todling
```

```
NAME

 atmens_prepobsens.csh  - prepare environment for ensemble-FSO calculations

SYNOPSIS

 atmens_prepobsens.csh  expid nymd nhms taub aver action

where
  expid  -  usual experiment name, e.g., b541iau
  nymd   -  date of analysis, as in YYYYMMDD
  nhms   -  time of analysis, as in HHMMSS
  taub   -  forecast interval EFSO calcualted for (in hours)
  aver   -  verification type (ana,asm,or niana)
  action -  the following are valid:
            setrc  - to generate acquire file with all needed for EFSO
            null   - to actually create all links and resolution conversions

DESCRIPTION

   This procedure is responsible for converting the output of the
 EnKF backward integration into ODS and producing whetever other
 diagnostic and statistic desired from the EFSO procedure.

 Example of valid command line:
 atmens_prepobsens.csh b541iau 20091019 000000 24 asm setrc

REQUIRED ENVIRONMENT VARIABLES

   AENSTAT_MPIRUN   - mp_stats MPI command line
   ATMENSETC        - location of experiment
   FVWORK           - location of work directory

OPTIONAL ENVIRONMENT VARIABLES

   ATMENS_FSO_JGRAD  - 1: use GMAO normalized error instead of EnKF internal norm
                       (default: 0)
   ATMENS_FSO_AVRFY  - 0: use central analysis/asm for verification
                       1: use non-inflated ensemble analysis for verification
                       (default: 0)
   ATMENS_FSO_MFCST  - 0: use central forecast for error definition
                       1: use ensemble mean forecast for error definition
                       2: use adj-derived sensitivity from central
                       (default: 0)
   NCSUFFIX          - suffix of hdf/netcdf files (default: nc4)
   DATADIR           - location where original data resides
                       (default: /archive/u/user)
   SRCEXPID          - original experiment (use when other than expid)
                       (default: expid)
```

```
SEE ALSO
  ut_prepobsens.csh - unit tester for this procedure

AUTHOR
  Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    Created modified: 01Apr2017   by: R. Todling
    Last modified: 16Apr2017      by: R. Todling
```

```
NAME

 atmens_recenter.csh  - recenter ensemble around desired mean

SYNOPSIS

 atmens_recenter.csh  expid nymd nhms ftype1 ftype2 ensloc cenloc infloc

 where
   expid  -  usual experiment name, e.g., u000_c92
   nymd   -  date of analysis, as in YYYYMMDD
   nhms   -  time of analysis, as in HHMMSS
   ftype1 -  member file type, e.g,, inc.eta
   ftype2 -  mean   file type, e.g,, ana.eta
   ensloc -  location of ensemble members and mean
   cenloc -  location of centering mean
   infloc -  location of perturbations for additive inflation

DESCRIPTION:

Example of valid command line:
atmens_recenter.csh u000\_c92 20091019 000000 inc.eta ana.eta \
        /discover/nobackup/rtodling/u000\_c92/atmens \
        /archive/u/rtodling/b541iau/ana/Y2009/M10

REQUIRED ENVIRONMENT VARIABLES

   ATMENSETC     - location of ensemble RC files
   ENSWORK       - location of work directory
   FVROOT        - location of DAS build

OPTIONAL ENVIRONMENT VARIABLES

   ADDINF_FACTOR      - inflation factor (such as 0.25)
   AENS_ADDINFLATION  - when set, apply additive inflation to
                          analyzed members
   AENS_DONORECENTER  - allow bypassing recentering
   AENS_RECENTER_DSTJOB  - distribute multiple works within
                             smaller jobs
   AENS_RECENTER_DSTJOB  - distribute multiple works within smaller jobs
   ASYNBKG            - background frequency (when adaptive inflation on)
   CENTRAL_BLEND      - 0 or 1=to blend members with central (def: 1)
   FVHOME             - location of experiment
   NCSUFFIX           - suffix of hdf/netcdf files
   ENSPARALLEL        - when set, runs all ensemble components in parallel
                        (default: off)
   ENSRECENTER_NCPUS  - when parallel ens on, this sets NCPUS for
                        recentering
                        (NOTE: required when ENSPARALLEL is on)
   RECENTER_WALLCLOCK - wall clock time to run dyn_recenter
```

```
                          (default 0:10:00)
   RECENTER_QNAME     - name of queue
                          (default: NULL, that is, let pbs pick)

REMARKS

 WARNING: This will overwrite each ensemble member file

AUTHOR
  Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    Last modified: 08Apr2013      by: R. Todling
```

```
NAME

 atmens_restarts.csh  - entry point to ensemble AOD analysis

SYNOPSIS

 atmens_restarts.csh  odir oexpid ndir nexpid nymdhh

where
  odir    -  original experiment location
               (e.g., archive location)
  oexpid  -  original experiment name
  ndir    -  new experiment location
               (i.e., nobackup area)
  nexpid  -  new experiment name
  nymdhh -  date of analysis, as in YYYYMMDDHH

DESCRIPTION

   This procedures copies all files needed to start an ensemble
  experiment from an existing ensemble experiment.

 Example of valid command line:
 atmens_restarts.csh /archive/u/rtodling e512a_rt \
        /discover/nobackup/rtodling e513T_rt 2014071709

REQUIRED ENVIRONMENT VARIABLES

OPTIONAL ENVIRONMENT VARIABLES

RESOURCE FILES

SEE ALSO

AUTHOR
  Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    Last modified: 15Oct2014      by: R. Todling
```

```
NAME
atmens_seasonal_dates.csh  - generate random dates
            falling within season of current date.

SYNOPSIS

 atmens_seasonal_dates.csh  expid nymd nhms

 where
   nymd   -  initial date of ensemble forecast, as in YYYYMMDD
   nhms   -  initial time of ensemble forecast, as HHMMSS
   nmem   -  number of members (random dates to generate)

DESCRIPTION

  This procedure generates random random dates from dates within
  database of NMC-perturbations falling within the season of the
  date specified at command line.

Example of valid command line:
 atmens_seasonal_dates.csh 20110325 12000 32

REQUIRED ENVIRONMENT VARIABLES

   ATMENSETC     - location of EnKF resource files
   FVROOT        - location of DAS build

OPTIONAL ENVIRONMENT VARIABLES

   ATMENS_VERBOSE - turn on shell verbose (default: 0)
   VERBOSE        - minimal echo of results (default: 1)

SEE ALSO
 randates.py           - generate random dates within given
                         range of dates (by J. Whitaker)
 ut_seasonal_dates.csh - unity tested for the present script

AUTHOR
  Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    Last modified: 08Apr2013      by: R. Todling
```

```
NAME

 atmens_stats.csh  - calculates statistics from ensemble members

SYNOPSIS

 atmens_stats.csh nmem ftype ensloc nymd nhms

where
  nmem   -  number of members to be created
  ftype  -  file type (e.g., bkg.eta, bkg.sfc, ana.eta)
  ensloc -  location to place generated ensemble
  nymd   -  date of members to calc stats for (as YYYYMMDD)
  nhms   -  time of members to calc stats for (as HHMMSS)

DESCRIPTION

 This script calculates statistics such as mean, RMS and
 energy-based spread from ensemble members. Two ways of
 performing these calculations are available. These are
 driven by the present (or absence) of the resource
 file mp_stats.rc, which triggers the most efficient way
 of performing all required calculations using a single
 pass through the input data. Alternatively, a series
 of calculations is performed where: (i) the mean of
 the members is computed, followed by removal of mean
 from members, followed by (iii) calculation of RMS;
 and finally performing (iv) calculation of
 energy-based spread.

 Example of valid command line:
  atmens_stats.csh 10 /archive/u/rtodling/u000_c72/atmens 20111201 210000

REQUIRED RESOURCE FILES

 mp_stats.rc              - needed when calculations use mp_stats.x
 atmens_incenergy.rc.tmpl - needed when energy-spread uses pertenergy.x

REQUIRED ENVIRONMENT VARIABLES

   AENSTAT_MPIRUN - command line for exec of mp_stats.x
   ASYNBKG    - frequency of background (minutes)
   EXPID      - experiment name
   FVROOT     - location of DAS build
   FVHOME     - location of experiment

OPTIONAL ENVIRONMENT VARIABLES

   ATMENSETC     - specify to provide location of pert-energy RC file
   ENSPARALLEL   - when set, runs all ensemble components in parallel
```

```
                    (default: off)
   AENSTAT_NCPUS - number of cpus to use for this procedure
                   (NOTE: required when ENSPARALLEL is on)
SEE ALSO
 mp_stats.x   - program to calculate statistics from fields in SDF files
 dyn_diff.x   - program to calculate difference between dyn-vector files
 GFIO_mean.x  - program to calculate averages from fields in SDF files
 pertenergy.x - calculates energy-based error (squared-difference)
 ut_atmens_stats.csh - unity-tester for this script (also helps produce
                       statistics when running offline).

AUTHOR
  Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    Last modified: 08Apr2013      by: R. Todling
```

```
NAME

 atmens_vtrack.csh  - calculate TC track for each member of ensemble

SYNOPSIS

 atmens_vtrack.csh  expid nymd nhms vtrkfrq

where
  expid     -  usual experiment name, e.g., b541iau
  nymd      -  starting date of forecast (as in YYYYMMDD)
  nhms      -  starting time of forecast (as in HHMMSS)

DESCRIPTION

 Example of valid command line:
 atmens_vtrack.csh b541iau 20091019 000000

REQUIRED RESOURCE FILES

  vtrack.rc       - controls track program
  vtrack.ctl.tmpl - grads template for tracking program
  vtx.ctl.tmpl    - grads template for relocator program

REQUIRED ENVIRONMENT VARIABLES

   ATMENSETC       - location of resource files
   FVHOME          - location of experiment
   FVROOT          - location of DAS build
   FVWORK          - location of work directory
   GID             - group id for PBS jobs
   VAROFFSET       - offset min from first synoptic time
   VTRKFRQF        - frequency of tracking calculation
   VTXLEVS         - pressure levels track works with

OPTIONAL ENVIRONMENT VARIABLES

   ATMENSLOC        - location of ensemble (default: FVHOME)
   NCSUFFIX         - suffix of hdf/netcdf files (default: nc4)
   ENSPARALLEL      - when set, runs all ensemble components
                         in parallel
                          (default: off)
   AENS_VTRACK_DSTJOB- distribute multiple works within smaller jobs
   ENSVTRK_NCPUS     - number of cpus used to run tracker
   VTRACK_WALLCLOCK  - wall clock time to run vtrack, default 1:00:00
   VTRACK_QNAME      - name of queue
                         (default: NULL, that is, let pbs pick)
   STRICT            - sets whether or not tcvitial must be present

SEE ALSO
```

```
  vtrack  - main vortex track driver

AUTHOR
  Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    Last modified: 01Sep2015      by: R. Todling
```

```
NAME

 atmos_eana.csh  - call desired Ensemble Analysis strategy.

SYNOPSIS

 atmos_eana.csh  expid nymd nhms

where
  expid  -  usual experiment name, e.g., b541iau
  nymd   -  date of analysis, as in YYYYMMDD
  nhms   -  time of analysis, as in HHMMSS

DESCRIPTION

 Current options for ensemble analyses strategies include:
       EnKF, EnGSI, and Simplified Ensemble.
 Only the EnKF and the Simplified Ensemble have been thoroughly
 tested and evaluated. The triggers in each case are determined
 depending on the presence of the following files in the ATMENSETC
 directory:
   EnKF:  obs1gsi_mean.rc
          obs1gsi_member.rc
          atmos_enkf.nml.tmpl
   EnGSI: gsi_mean.rc
          gsi_member.rc
   Simplified Ensemble: easyeana.rc

Example of valid command line:
atmos_eana.csh b541iau 20091019 000000

REQUIRED ENVIRONMENT VARIABLES

   ATMENSETC     - location of ensemble resource files
   FVHOME        - location of experiment
   FVROOT        - location of DAS build
   FVWORK        - location of work directory

SEE ALSO

  atmos_enkf.csh - driver for EnKF analysis
  atmos_egsi.csh - driver for ensemble of GSI analysis
  atmos_eezy.csh - driver for simplified scheme

AUTHOR
  Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    Last modified: 08Apr2013      by: R. Todling
```

```
NAME

 atmos_eaod.csh  - entry point to ensemble AOD analysis

SYNOPSIS

 atmos_eaod.csh  expid nymd nhms freq

where
  expid  -  usual experiment name, e.g., b541iau
  nymd   -  date of analysis, as in YYYYMMDD
  nhms   -  time of analysis, as in HHMMSS
  freq   -  frequency of AOD analysis, as in HHMMSS

DESCRIPTION

  This procedures handles the analysis of aerosols in the ensembe DAS.
  In its simplest form this procedure makes the central DAS AOD
  analysis available to each of the members of the ensemble.

 Other options, controled by the env variable AENS_GAAS_OPT (see below)
 include the possibility of running the PSAS-based AOD analysis for
 each member of the ensemble.

 Additionally, or alternatively, this procedure also allows for using
 the EnSRF to analyze aerosols through an ensemble-based approach.

 When using the EnSRF, the option exists to analyze either AOD or the
 full 3D concentration fields. These are controlled directly by the
 parameters in the EnKF namelist file (not by.  this script); this
 script is, however, capable of automatically recongnizing between
 the two options and making adequate decisions from there.

 Example of valid command line:
 atmos_eaod.csh b541iau 20091019 000000 030000

REQUIRED ENVIRONMENT VARIABLES

 ATMENSETC     - location of ensemble RC files
 ATMENSLOC     - location of ensemble members
 FVHOME        - location of experiment
 FVROOT        - location of DAS build
 FVWORK        - location of work directory

OPTIONAL ENVIRONMENT VARIABLES

 AENKFAERO_NCPUS   - when parallel ens on, this sets NCPUS for AGCM
                     integration
 ATMENKFAERO_WALLCLOCK - wall-clock time to run EnKF, default 1:00:00
 ATMENKFAERO_QNAME - name of queue (default: NULL, that is, let pbs pick)
```

```
 ENSACQ_WALLCLOCK  - wallclock time for acquire job
                     (default: 2:00:00)
 ENSPARALLEL       - when set, runs all ensemble components in parallel
 GAAS_ANA          - triggers use of AOD analysis
                     (default: 0; no ANA)
 AENS_GAAS_OPT     - set options for AOD ensemble analysis:
                     1 -> use central analysis for all members
                     2 -> run GAAS analysis for each member
                     3 -> do (2), and EnKF-based AOD analysis
                          (off aod.or.concentrations)
                     4 -> EnKF-based AOD analysis
                          (off aod.or.concentrations)
                    (default: 1)
 AERO_FROM_ENKF    - when specified will replace PSAS analysis with
                     those from EnKF
 NCSUFFIX          - suffix of hdf/netcdf files
                     (default: nc4)

RESOURCE FILES

 gmao_aero_hybens_info.xNLATyNLONlNLEV.rc - opt rc allowing definition
                                            of horizontal and vertical
                                            EnKF localization
                                            (DEFAULT: see enkf.nml)

SEE ALSO

  atmos_enkf.csh, run_gaas_ana.csh, calcaod.csh

AUTHOR
  Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    In collaboration with: V. Buchard
    Last modified: 26Mar2017      by: R. Todling
```

```
NAME

 atmos_eezy.csh  - create ensemble from central analysis

SYNOPSIS

 atmos_eezy.csh  expid nymd nhms

where
  expid  -  usual experiment name, e.g., b541iau
  nymd   -  date of analysis, as in YYYYMMDD
  nhms   -  time of analysis, as in HHMMSS

DESCRIPTION

  This is the driver for the Simplified Ensemble Scheme. In
this procedure, an ensemble of analysis is created by simply
perturbing the central analysis with scaled NMC-like 48-24 hour
perturbations. The resolution of the ensemble is defined in the
resource file easyeana.rc.

Example of valid command line:
atmos_eezy.csh b541iau 20091019 000000

REQUIRED RESOURCE FILES

 easyeana.rc - specified parameters to simplified scheme

REQUIRED ENVIRONMENT VARIABLES

   ATMENSETC      - location of ensemble resource files
   ATMENSLOC      - location of ensemble, usually FVHOME
   FVHOME         - location of experiment
   FVROOT         - location of DAS build
   FVWORK         - location of work directory
   STAGE4HYBGSI   - location of where central analysis resides

REMARKS

 1. This option is triggered by the presence of easyeana.rc in ATMENSETC

SEE ALSO
  atmos_eana.csh - entry-point of ensemble analysis

AUTHOR
  Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    Last modified: 08Apr2013      by: R. Todling
```

```
NAME

 atmos_egsi.csh  - run ensemble of GSI analysis

SYNOPSIS

 atmos_egsi.csh  expid nymd nhms

where
  expid  -  usual experiment name, e.g., b541iau
  nymd   -  date of analysis, as in YYYYMMDD
  nhms   -  time of analysis, as in HHMMSS

DESCRIPTION

 This procedures generates an ensemble of analysis by running GSI
 for each individual set of member-backgrounds.

Example of valid command line:
atmos_egsi.csh b541iau 20091019 000000

REQUIRED RESOURCE FILES

 gsi_mean.rc   - specified parameters to mean observer/analysis
 gsi_member.rc - specified parameters to member observer/analysis

REQUIRED ENVIRONMENT VARIABLES

   FVHOME         - location of experiment
   FVROOT         - location of DAS build
   FVWORK         - location of work directory

REMARKS
  1) This procedure has not been properly evaluated yet.

SEE ALSO
  atmos_eana.csh - entry-point of ensemble analysis

AUTHOR
  Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    Last modified: 08Apr2013      by: R. Todling
```

```
NAME

 atmos_enkf.csh  - run atmospheric EnKF analysis

SYNOPSIS

 atmos_enkf.csh  expid nymd nhms

where
  expid  -  usual experiment name, e.g., b541iau
  nymd   -  date of analysis, as in YYYYMMDD
  nhms   -  time of analysis, as in HHMMSS

DESCRIPTION

 This procedure creates an ensemble of analysis by running
 J. Whitaker ensemble Kalman filter (EnKF). This is the same
 software-basis as that presently used at NCEP 3DVAR-Hybrid.
 The parameter settings used for GEOS are specific to GEOS,
 including resoltion, and vertical and horizontal localizations.

 Example of valid command line:

 atmos_enkf.csh b541iau 20091019 000000

REQUIRED RESOURCE FILES

 atmos_enkf.nml.tmpl - sets parameters for EnKF software

REQUIRED ENVIRONMENT VARIABLES

 ATMENSETC      - location of EnKF resource files
 FVHOME         - location of experiment
 FVROOT         - location of DAS build
 FVWORK         - location of work directory
 MPIRUN_ATMENKF - define mpi command for GSIsa.x

OPTIONAL ENVIRONMENT VARIABLES

 NCSUFFIX          - suffix of hdf/netcdf files
                     (default: nc4)
 ENSPARALLEL       - when set, runs all ensemble components in parallel
                     (default: off)
 AENKF_NCPUS       - when parallel ens on, this sets NCPUS for AGCM
                     integration
 ATMENKF_WALLCLOCK - wall clock time to run EnKF
                       (default 1:00:00)
 ATMENKF_QNAME     - name of queue
                       (default: NULL, that is, let pbs pick)
```

```
SEE ALSO

  atmos_eana.csh - entry-point of ensemble analysis

AUTHOR
  Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    Last modified: 03Mar2017      by: R. Todling
```

```
NAME

 atmos_ens2gcm.csh  - construct IAU increment for each member

SYNOPSIS

 atmos_ens2gcm.csh  expid nymd nhms

where
  expid  -  usual experiment name, e.g., b541iau
  nymd   -  current analysis date
  time   -  current analysis time

DESCRIPTION

 This procedure constructs IAU increment for each member
 of the ensemble from corresponding analysis and background
 fields.

 Example of valid command line:
 atmos_ens2gcm.csh b541iau 20091019 000000

REQUIRED ENVIRONMENT VARIABLES

  ATMENSETC       - location of pertinent resource files
  FVHOME          - location of experiment
  FVROOT          - location of DAS build
  FVWORK          - location of work directory
  GID             - group id to run job under

OPTIONAL ENVIRONMENT VARIABLES

  AENS_IAU_DSTJOB- distribute multiple works within smaller jobs
  NCSUFFIX       - suffix of hdf/netcdf files (default: nc4)
  ENSPARALLEL    - when set, runs all ensemble components in parallel
                   (default: off)
  ENSIAU_NCPUS   - when parallel ens on, this sets NCPUS for IAU calculation
  IAU_WALLCLOCK  - wall clock time to run makeiau, default 0:10:00
  IAU_QNAME      - name of queue (default: NULL, that is, let pbs pick)
  MPIRUN_ENSIAU  - specifies mprun command line, needed when ENSPARALLEL is on

SEE ALSO

AUTHOR
  Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    Last modified: 08Apr2013      by: R. Todling
```

```
NAME

 edas_scheduler.csh  - scheduler for running EnADAS in parallel
            with central ADAS

SYNOPSIS

 edas_scheduler.csh

DESCRIPTION

 This allows launching the ensemble analysis as soon as the
 central analysis is available. The aim of the scheduler is
 to expedite the cycle and maximize efficiency when running
 the hybrid-variational DAS.

 The trigger for this mode of execution is the presence of
 the file edas_scheduler.j in the run directory of the
 experiment and the environment setting
     setenv ENSPARALLEL 2
 in the AtmEnsConfig.csh configuration script.

REQUIRED ENVIRONMENT VARIABLES

   ATMENSETC - location of ensemble resource files
   DDASJNAME - name of script controling central DAS (e.g., g5das.j)
   EXPID     - experiment name
   FVHOME    - location of experiment
   FVROOT    - location of DAS build
   FVWDIR    - root location of fvwork/enswork
   VAROFFSET - offset time from first synoptic time (min)

REMARKS

 1. Only very limited testing has been done with this mode of
    execution. There are still known issues related to the
    scheduling, thus this is not the presently recommended mode
    of running.

SEE ALSO

 atm_ens.j        - job script, driver of the Ensemble Analysis
 edas_scheduler.j - job script, driver of the Scheduler

AUTHOR
  Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    Last modified: 18Jan2015      by: R. Todling
```

```
NAME

 gcm_ensemble.csh  - run multiple copies of (atmospheric) GCM

SYNOPSIS

 gcm_ensemble.csh  expid nymd nhms tfcst nlons nlats

where
  expid  -  usual experiment name, e.g., b541iau
  nymd   -  initial date of forecast, as in YYYYMMDD
  time   -  initial time of forecast, as HHMMSS
  tfcst  -  forecast length in hours
  nlons  -  number of longitudes in bkg (im) (for history)
  nlats  -  number of latitudes  in bkg (jm) (for history)

DESCRIPTION

  This procedure runs multiple copies of the, presently
  atmospheric, GCM. These integrations are, in principle,
  forced with an ensemble of IAU-increments).

  The usual, main RC files for the GCM must be present in
  the directory defined by ATMENSETC. Typically, this are
  the files listed under REQUIRED RESOUCE FILES below.

  Furthermore, a file named lnbcs_ens must be placed under
  the run directory. The reason why this file is not placed
  under ATMENSETC is simply because executables are to live
  either in the bin of the build or in the run directory, and
  lnbcs is an executable script.

  Just as in the regular DAS, similar tricks to bootstrap restarts
  apply here for the ensemble of GCMs, that is, restarts can be boot-
  strapped by placing a file named AGCM.BOOTSTRAP.rc.tmpl under the
  ATMENSETC directory. Analogously, once can request the GCM to
  integrate beyond the usual 12-hour period per cycle by placing
  specific files to control those particular instances. For example,
  if one wishes to integrate the 00z predictions out to 24-hours,
  placing files CAP_21.rc.tmpl and HISTAENS_21.rc.tmpl in the
  ATMENSETC directory with properly defined length of integration
  and desirable history will do it.

Example of valid command line:
gcm_ensemble.csh b541iau 20091018 210000 12 144 91

REQUIRED RESOURCE FILES

  CAP.rc.tmpl      - determine length of integration
  AGCM.rc.tmpl     - defines specific restarts, and GCM parameters
```

```
  HISTAENS.rc.tmpl - defines output of ensemble of GCMs

REQUIRED ENVIRONMENT VARIABLES

   ATMENSETC     - location of resource files
   ATMENSLOC     - location of current ensemble
   ASYNBKG       - frequency of background (minutes)
   FVBCS         - location of fvInput
   FVHOME        - location of experiment
   FVROOT        - location of DAS build
   FVWORK        - location of work directory
   GID           - group ID to run job under
   MPIRUN_ENSGCM - define mpi command for GEOSgcm.x
   RSTSTAGE4AENS - location of restarts
   TIMEINC       - analysis frequency (minutes)

OPTIONAL ENVIRONMENT VARIABLES

   ATMGEPS        - trigger for GEOS EPS
   NCSUFFIX       - suffix of hdf/netcdf files (default: nc4)
   ENSPARALLEL    - when set, runs all ensemble components in parallel
                    (default: off)
   ENSGCM_NCPUS   - when parallel ens on, this sets NCPUS for AGCM
                    integration
   AENS_GCM_DSTJOB- distribute multiple works within smaller jobs
   AGCM_WALLCLOCK - wall clock time to run agcm
                      (default 1:00:00)
   AGCM_QNAME     - name of queue
                    (default: NULL, that is, let pbs pick)
   ATMENS_DO4DIAU - trigger to run 4DIAU

REMARKS

  1. When atmens_rst_regrid.rc is present in ATMENSETC
     this script will regrid the restart in that file
     to the desired resolution; these will not be recycled.
  2. When running GEOS EPS the length of forecast and history
     can be controlled by dropping the following two files:
         CAP_hh.rc.tmpl (or simply CAP.rc.tmpl)
         HISTAGEPS_hh.rc.tmpl (or simply HISTAGEPS.rc.tmpl)
     in the FVHOME/ageps directory.

SEE ALSO
  atmos_ens2gcm.csh - calculation of IAU increments
  atm_ens_geps.j    - main job script controlling GEPS

AUTHOR
  Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    Last modified: 20Apr2017      by: R. Todling
```

```
NAME

 gen_ensbkg.csh  - artificially generates ensemble backgrounds

SYNOPSIS

 gen_ensbkg.csh  BKGLOC ENSLOC expid nmem

where
  BKGLOC - archive location where original experiment resides,
           e.g., /archive/u/rtodling/
  ENSLOC -  location to place generated ensemble
  expid  -  usual experiment name, e.g., b541iau
  nmem   -  number of members to be created

DESCRIPTION

 Articificially generates ensemble backgrounds
 by adding NMC-like perturbations to the initial
 central backgrounds

 Example of valid command line:
 gen_ensbkg.csh /discover/nobackup/rtodling/myexp/recycle \
         /discover/nobackup/rtodling/myexp/atmens 10

REQUIRED ENVIRONMENT VARIABLES

   ATMENSETC - location of ensemble resource files
   ASYNBKG   - frequency of background files (minutes)
   FVHOME    - location of experiment
   FVROOT    - location of DAS build

OPTIONAL ENVIRONMENT VARIABLES

   ATMENS_DOSTATS    - allows bypass calculation of statistics
                       (default: 1 - do it)
   ADDINF_FACTOR      - inflation factor (such as 0.25)
   CENTRAL_BLEND      - 0 or 1=to blend members with central (def: 1)

REMAKRS
 1. This procedure is largely obsolete

SEE ALSO
  gen_ensrst.csh

AUTHOR
  Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    Last modified: 08Apr2013      by: R. Todling
```

```
NAME

 gen_ensemble.csh  - artificially generates ensemble from
                     an existing experiment

SYNOPSIS

 gen_ensemble.csh  EXPLOC expid inymd inhms newid enymd enhms nmem ensdir

where
  EXPLOC - archive location where original experiment resides,
           e.g., /archive/u/rtodling/
  expid  -  usual experiment name, e.g., b541iau
  inymd  -  date of files in archive to serve as samples, as in YYYYMMDD
            i.e., ensemble will be generated from file centered
            around this date
  inhms  -  time of files in archive, as in HHMMSS
  newid  -  new experiment id
  enymd  -  actual date user wants ensemble at, as in YYYYMMDD
  enhms  -  actual time user wants ensemble at, as in HHMMSS
  nmem   -  number of members to be created
  ensdir -  location to place generated ensemble

DESCRIPTION
   This procedure creates an ensemble of background by taking backgrounds
 from different dates of an existing experiment and pretending them to be
 all valid at a given date/time.
   The procedure also creates an ensemble of GCM restarts by copying
 restarts from an existing experiment, at a given time, and
 re-time-tagging them to pretend they all fall on the desired
 initial date/time of the experiment to be performed.

Example of valid command line:
gen_ensemble.csh /discover/nobackup/rtodling/expid/recycle b541iau \
        20091019 000000 z000_b72 20090801 000000 10 \
        /discover/nobackup/rtodling/Gen_Ens

REQUIRED ENVIRONMENT VARIABLES

   ASYNBKG   - frequency of background (minutes)
   FVROOT    - location of DAS build
   TIMEINC   - frequency of analysis (minutes)

OPTIONAL ENVIRONMENT VARIABLES

   ATMENS_DOSTATS    - allows bypass calculation of statistics
                       (default: 1 - do it)
   SIMULATE_ENSEMBLE - dry test: exec without actual copying

REMARKS
```

1. After multiple versions of the ensemble initialization procedure we have now settle on something rather simpler than what performed here. The present possible modes of initializing the ensemble are more effective, and less prone to spind-up/down issues.
2. Therefore, this procedure is largely obsolete, though its sub-components are stil useful.

SEE ALSO

gen_ensbkg.csh - generates ensemble of backgrounds
gen_ensrst.csh - generates ensemble of restarts

AUTHOR

Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
Last modified: 08Apr2013 by: R. Todling

```
NAME

 gen_ensrst.csh  - artificially generates ensemble of GCM restarts

SYNOPSIS
 gen_ensrst.csh  nymd nhms oexpid nexpid rstloc ensloc

where
  nymd    -  initial date of experiment, as in YYYYMMDD
  nhms    -  initial time of experiment, as in HHMMSS
  oexpid  -  name of experiment for original set of restarts
  nexpid  -  name of present (new) hyrbrid experiment
  rstloc  -  location of original restarts
  ensloc  -  location of initial ensemble for present experiment

DESCRIPTION

 Articificially generates ensemble of GCM (restarts) initial
 conditions. This is done by simply associating the ens of restarts
 to a pre-existing set of restarts. User must be mindful of two
 situations:
    (i) full resolution hybrid: initial rsts can be associated to
        those under recycle
   (ii) dual resolution hybrid: an extra set of restarts at reduced
        resolution must be made available for the same initial time
        of the experiment (that is, as the restarts under recycle)

 Example of valid command line:
 gen_ensrst.csh 20090731 210000 z000_b72 u000_c72 \
         /discover/nobackup/rtodling/z000_b72/recycle \
         /discover/nobackup/rtodling/u000_c72/atmens

REQUIRED ENVIRONMENT VARIABLES

   FVROOT       - location of DAS build

OPTIONAL ENVIRONMENT VARIABLES

   NCSUFFIX     - suffix of bkg files (default: nc4)
   RECENTERBKG  - force recentering of bkg.eta files
                  (default: 0; not)

SEE ALSO
  atmens_recenter.csh - control recentering of ensemble members

AUTHOR
  Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    Last modified: 08Apr2013      by: R. Todling
```

```
NAME

 gen_nmcperts.csh  - generate NMC-perturbations from (OPS) forecasts

SYNOPSIS

 gen_nmcperts.csh  fcstdir workdir expid expout nymd nhms ndays

where
  fcstdir - path of experiment holding forecasts
  workdir - path of directory where to write NMC perturbations to
  expid  -  name of experiment holding forecasts
  expout -  name of NMC perturbation
  nymd   -  date of initial forecast, as in YYYYMMDD
  time   -  time of intial forecast, as HHMMSS
  ndays  -  number of days to generate pert from current date/time

DESCRIPTION

  This script looks in the archive and generates the 48-24-hr NMC-like
  perturbations making up the database used in the additive inflation
  strategy of the ensemble DAS.

Example of valid command line:
gen_nmcperts.csh /archive/u/dao_ops/GEOS-5.7.2/GEOSadas-5_7_2_p5_m1 \
        /discover/nobackup/rtodling/NMCperts \
        e572p5_fp e572_fp 20120523 000000 1

REQUIRED ENVIRONMENT VARIABLE

   FVROOT        - location of DAS build

OPTIONAL ENVIRONMENT VARIABLE

   ATMENS_VERBOSE - set verbose on
   NCSUFFIX       - suffix for SDF files (default: nc4)

SEE ALSO
  ut_nmcperts.j - unit tester (helps generate perts when needed)

AUTHOR
  Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    Last modified: 08Apr2013      by: R. Todling
```

```
NAME

 jobmonitor.csh  - monitor events of a parallel section

SYNOPSIS

 jobmonitor.csh  nevents calledby workdir yyyymmddhh

 where
  nevents    -  number of parallel events
                 (i.e., no. of ensemble members)
  calledby   -  name of calling program
  workdir    -  location where work takes place
  yyyymmddhh -  date/hour of events

DESCRIPTION

   This procedure monitors the completion of parallel procedures
   running the within atmopheric ensemble DAS.

OPTIONAL ENVIRONMENT VARIABLES

  JOBMONITOR_DELSLEEP_SEC - specify time interval to check job
                            termination (seconds)
                              (default: 20)
  JOBMONITOR_MAXSLEEP_MIN - specify max allowed time per
                            procedure (minutes); procedure
                            means e.g. all of observer calls

AUTHOR
  Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    Last modified: 08Apr2013      by: R. Todling
```

```
NAME

 makeiau.csh  - create IAU increment for given member

SYNOPSIS

 makeiau.csh  expid member nymd nhms

where
  expid  -  usual experiment name, e.g., b541iau
  member -  number of member to operate on
  nymd   -  date of analysis, as in YYYYMMDD
  nhms   -  time of analysis, as in HHMMSS

REQUIRED RESOURCE FILES

 mkiau.rc.tmpl - when running cubed, this must be under ATMENSETC

REQUIRED ENVIRONMENT VARIABLES

   ATMENSETC - location of ensemble RC files
   FVHOME    - location of experiment
   FVROOT    - location of DAS build
   MPIRUN_ENSIAU - mpi-run command to handle executable

OPTIONAL ENVIRONMENT VARIABLES

   NCSUFFIX  - suffix for netcdf files (default: nc4)

REMARKS
 1. This procedure is largely obsolete.

SEE ALSO

  atmos_ens2gcm.csh - driver of IAU-increment calculation

AUTHOR
  Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    Last modified: 08Apr2013      by: R. Todling
```

```
NAME

 obsvr_ensemble.csh  - run observer over mean background and
                       each ensemble member

SYNOPSIS

 obsvr_ensemble.csh  obsclass expid nymd nhms

where
  obsclass  -  observation class
  expid     -  usual experiment name, e.g., b541iau
  nymd      -  date of analysis, as in YYYYMMDD
  nhms      -  time of analysis, as in HHMMSS

DESCRIPTION
   This procedure implements a driver for the ensemble of observers
 required to run the EnKF and some other variants of the Ensemble DAS.
 This procedure is based on the central DAS acquiring of observations,
 and calling of the GSI.

 Example of valid command line:
 obsvr_ensemble.csh ncep_prep_bufr,ncep_1bamua_bufr \
                    b541iau 20091019 000000

REQUIRED RESOURCE FILES

 Depending on the ensemble strategy under consideration:
   obs1gsi_mean.rc   - specify observer parameter for mean observer
   obs1gsi_member.rc - specify observer parameter for member observer
   gsi_mean.rc       - similar to obs1gsi_mean.rc, but also
                           runs minimization
   gsi_member.rc     - similar to obs1gsi_member.rc, but also
                           runs minimization

REQUIRED ENVIRONMENT VARIABLES

   ATMENSETC       - location of resource files
   ATMENSLOC       - location of atmos-ensemble
   FVHOME          - location of experiment
   FVROOT          - location of DAS build
   FVWORK          - location of work directory
   GID             - group id to run job under
   MPIRUN_ENSANA   - define mpi command for GSIsa.x
   TIMEINC         - analysis frequency (minutes)
   VAROFFSET       - offset time from initial synoptic time (minutes)

OPTIONAL ENVIRONMENT VARIABLES

   NCSUFFIX         - suffix of hdf/netcdf files (default: nc4)
```

```
   ENSMEANONLY      - run observer for ensemble mean only
   ENSPARALLEL      - when set, runs all ensemble components in parallel
                      (default: off)
   ENSGSI_NCPUS     - when parallel ens on, this sets NCPUS for
                       Observer calculation
   AENS_OBSVR_DSTJOB- distribute multiple works within smaller jobs
   OBSVR_WALLCLOCK  - wall clock time to run observer, default 1:00:00
   OBSVR_QNAME      - name of queue
                       (default: NULL, that is, let pbs pick)

SEE ALSO

  setobsvr.csh - prepare observations to be used by
                 observer (based on fvssi)
  gsidiags     - creates diag-files from GSI
                 output (as used in central DAS)

AUTHOR
  Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    Last modified: 08Apr2013      by: R. Todling
```

```
NAME

 obsvr_ensfinal.csh  - run observer over mean analysis to get OmAs

SYNOPSIS

 obsvr_ensfinal.csh  expid nymd nhms anadir

where
  expid     -  usual experiment name, e.g., b541iau
  nymd      -  date of analysis, as in YYYYMMDD
  nhms      -  time of analysis, as in HHMMSS
  anadir    -  directory holding ensemble analyses

DESCRIPTION
   This procedure runs the observer after the ensemble analysis
 to get an approximate estimate of the OmA's. Since the input
 fields to this final-pass observer correspond to an update
 of the mean backgrounds this is considered to give an approximate
 estimate only.

 Example of valid command line:
 obsvr_ensfinal.csh b541iau 20091019 000000 ENSWORK/updated_ens

REQUIRED RESOURCE FILES

 This program is triggered by the presence of the following
 resource files:
  obs1gsi_member.rc             - specify member observer parameters
  GSI_GridComp_ensfinal.rc.tmpl - specify parameter for final obsvr

REQUIRED ENVIRONMENT VARIABLES

   ATMENSETC       - location of resource files
   FVHOME          - location of experiment
   FVROOT          - location of DAS build
   FVWORK          - location of work directory
   GID             - group id to run job under
   MPIRUN_ENSANA   - define mpi command for GSIsa.x
   TIMEINC         - analysis frequency (minutes)
   VAROFFSET       - offset time from initial synoptic time (minutes)

OPTIONAL ENVIRONMENT VARIABLES

   NCSUFFIX         - suffix of hdf/netcdf files (default: nc4)
   ENSPARALLEL      - when set, runs all ensemble components in parallel
                      (default: off)
   ENSGSI_NCPUS     - when parallel ens on, this sets NCPUS for
                        Observer calculation
   OBSVR_WALLCLOCK  - wall clock time to run observer, default 1:00:00
```

```
   OBSVR_QNAME      - name of queue
                        (default: NULL, that is, let pbs pick)

SEE ALSO

  post_eana - the post-analysis will call this procedure

REMARKS

  post_eana - the post-analysis will call this procedure

AUTHOR
  Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    Last modified: 15Apr2013      by: R. Todling
```

```
NAME

 post_eana.csh  - post ensemble-analysis calculations

SYNOPSIS

 post_eana.csh  expid nymd nhms

where
  expid  -  usual experiment name, e.g., b541iau
  nymd   -  date of analysis, as in YYYYMMDD
  nhms   -  time of analysis, as in HHMMSS

DESCRIPTION

   This procedure is responsible for calculating the necessary (and
 desired) statistics from the ensemble of analysis, and recentering
 the ensemble about the central analysis after the ensemble mean
  analysis is available.

 Example of valid command line:
 post_eana.csh b541iau 20091019 000000

REQUIRED ENVIRONMENT VARIABLES

   FVHOME        - location of experiment
   FVROOT        - location of DAS build
   FVWORK        - location of work directory

OPTIONAL ENVIRONMENT VARIABLES

   NCSUFFIX      - suffix of hdf/netcdf files (default: nc4)
   ENSPARALLEL   - when set, runs all ensemble components in parallel
                   (default: off)
   AENKF_NCPUS   - when parallel ens on, this sets NCPUS for
                    AGCM integration

SEE ALSO
   atmens_stats.csh  - calculates required/desired statistics
                       from ensemble
atmens_recenter.csh  - recenters ensemble (and applies
                       additive inflation)
 obsvr_ensfinal.csh  - calculate OmAs from mean analysis
atmens_obsimp0hr.csh - calculate OmAs from mean analysis

AUTHOR
  Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    Last modified: 08Apr2013      by: R. Todling
```

```
NAME

 post_efso.csh  - post ensemble-FSO calculations

SYNOPSIS

 post_efso.csh  expid nymd nhms

where
  expid  -  usual experiment name, e.g., b541iau
  nymd   -  date of analysis, as in YYYYMMDD
  nhms   -  time of analysis, as in HHMMSS
  taub   -  forecast interval EFSO calcualted for (in hours)

DESCRIPTION

   This procedure is responsible for converting the output of the
 EnKF backward integration into ODS and producing whetever other
 diagnostic and statistic desired from the EFSO procedure.

 Example of valid command line:
 post_efso.csh b541iau 20091019 000000

REQUIRED ENVIRONMENT VARIABLES

   FVHOME        - location of experiment
   FVROOT        - location of DAS build
   FVWORK        - location of work directory

OPTIONAL ENVIRONMENT VARIABLES

   NCSUFFIX      - suffix of hdf/netcdf files (default: nc4)
   ENSPARALLEL   - when set, runs all ensemble
                   components in parallel
                   (default: off)

SEE ALSO
  atmens_obsimp0hr.csh - calculate OmAs from mean analysis

AUTHOR
  Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    Last modified: 13Apr2017      by: R. Todling
```

```
NAME

 post_egcm.csh  - post processing after ensemble of GCMs

SYNOPSIS

 post_egcm.csh  expid nymd nhms toffset ensloc

where
  expid   -  usual experiment name, e.g., b541iau
  nymd    -  initial date of forecast, as in YYYYMMDD
  nhms    -  initial time of forecast, as HHMMSS
  toffset -  time offset to start calculating stats (min)
  ensloc  -  location of ensemble members

DESCRIPTION

   This procedure will calculate ensemble statistics for
 each of the output streams present in the COLLECTIONS
 table of the HISTAENS.rc.tmpl file. When initial-time
 specific history exists, this will take precedence over
 HISTAENS.rc.tmpl, i.e., if a file HISTAENS_21.rc.tmpl
 is present in the ATMENSETC directory, this will be
 used to determined the output stream to work from instead.

 Alternatively, still, if not all output streams are to be
 worked on, the user may specify it's own subset of streams
 to calculate statistics for (you must have at least bkg.sfc
 and bkg.eta in this list). This can be done by placing a file
 named post_egcm.rc under ATMENSETC with a trimmed version of
 the COLLECTIONS table in the history RC. The same idea applies
 to this file for choices for different initial times, that is,
 it is also possible to have a files like post_egcm_21.rc.

 Example of valid command line:
 post_egcm.csh b541iau 20091018 210000 360 FVWORK/updated_ens

REQUIRED ENVIRONMENT VARIABLES

   ASYNBKG       - frequency of background (minutes)
   ATMENSETC     - location of ensemble RC files
   FVHOME        - location of experiment
   FVROOT        - location of DAS build
   TIMEINC       - analysis frequency (minutes)

OPTIONAL ENVIRONMENT VARIABLES

   NCSUFFIX       - suffix of hdf/netcdf files (default: nc4)

OPTIONAL RESOURCE FILES
```

```
 post_egcm.rc - user specific collection subset

SEE ALSO

   atmens_stats.csh - calculates required/desired statistics
                      from ensemble

AUTHOR
  Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    Last modified: 08Apr2013      by: R. Todling
```

```
NAME

 recenter.csh  - recenter ensemble of analyses

SYNOPSIS

 recenter.csh  expid member nymd hh ftype1 ftype2 CENLOC INFLOC

where
  expid  -  usual experiment name, e.g., b541iau
  member -  number of member to operate on
  nymd   -  date of analysis, as in YYYYMMDD
  hh     -  time of analysis, as in HH
  ftype1 -  type of input field (i.e, ana.eta or inc.eta)
  ftype2 -  type of central field (i.e, ana.eta)
  CENLOC -  location of central analysis
  INFLOC -  location of inflating perturbations
            (set to NONE when not applicable)

DESCRIPTION
   This script provides a wrapper for the call to dyn_recenter.x, which
 is the program actually responsible for recentering an ensemble member
 around a given (typically, central DAS) analysis.

   The dyn_recenter program is also responsible for applying additive
inflation to each member of the ensemble.

REQUIRED ENVIRONMENT VARIABLES

   ATMENSETC - location of ensemble resource files
   FVHOME    - location of experiment
   FVROOT    - location of DAS build

OPTIONAL ENVIRONMENT VARIABLES

  NOTE: this procedure cannot have optional env variables

SEE ALSO

 atmens_recenter.csh - driver script for recentering of members
 dyn_recenter.x      - program that actually recenters given member

AUTHOR
  Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    Last modified: 08Apr2013      by: R. Todling
```

```
NAME
     setobsvr - Set up observer (usually for running analysis)

SYNOPSIS

     setobsvr [-h] [-log] [-cprs]  fvhome  fvwork

DESCRIPTION

     This script prepares observation files so observer can run.

     On input:
     --------
     fvhome     experiment home directory,
                e.g., /scratch1/$user/v000_b55
     fvwork     working directory, e.g., $TMPDIR

     Restarts, namelists and resource files are  expected to
     be available in $FVHOME/recycle and $FVHOME/run.  The
     run is performed in $FVWORK, and the output files are
     left there for archival by the calling script.

OPTIONS

     -h           prints this page
     -strict      returns with non-zero error if acquire fails to
                  resolve all input files
     -obsclass    cls1,cls2...
                  observation data classes, such as
                  conv_tovs,ssmi_wentz_tpw
     -log         log warning and errors to file

ENVIRONMENT VARIABLES

  SPECRES     resolution of spectral backgrounds
                 (254 for T254, 62 for T62, etc)
  VAROFFSET   time offset (abs value) between analysis time and
                 initial time of ana window
  TIMEINC     analysis time window
                 (6-hr for 3dvar; 6,12,18,24,etc for 4dvar)


ENVIRONMENT VARIABLES (optional)

  STRICT      when set, will crash if obs files missing

SEE ALSO

    fvssisetup    Experiment setup utility
    fvssi/gsi.j   Main experiment script created by fvssisetup.
```

```
AUTHOR
    Based on fvssi initially coded by Carlos Cruz
    Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    Last modified: 08Apr2013      by: R. Todling
```

```
NAME

 setperts.csh  - set perturbations for additive inflation

SYNOPSIS

 setperts.csh  expid member nymd nhms freq INFLOC

where
  expid  -  usual experiment name, e.g., b541iau
  member -  number of member to operate on
  nymd   -  date of analysis, as in YYYYMMDD
  nhms   -  time of analysis, as in HHMMSS
  freq   -  frequency of perturbations, in minutes
  INFLOC -  location of inflating perturbations
            (set to NONE when not applicable)

DESCRIPTION
   This procedure fetches perturbations from the database
 making them available to the present ensemble cycle. This
 also invokes the necessary procedures to calculate and
 remove the mean from the perturbations.

REQUIRED ENVIRONMENT VARIABLES

  ATMENSETC - location of ensemble resource files
  ASYNBKG   - frequency of backgrounds (minutes)
  FVHOME    - location of experiment
  FVROOT    - location of DAS build
  FVWORK    - work directory

OPTIONAL ENVIRONMENT VARIABLES

  AENS_PERTS_DSTJOB- distribute multiple works
                     within smaller jobs
  ENSPARALLEL - performs calculation in parallel
                  (default: 0)
  NCSUFFIX    - suffix for netcdf files
                  (default: nc4)

SEE ALSO
  acquire_ensperts.csh - acquire (link) perturbation residing in database
  mp_stats.x           - provides one possible way to remove mean of perts

AUTHOR
  Amal El Akkraoui (Amal.ElAkkraoui@nasa.gov), NASA/GMAO
    Last modified: 08Apr2013      by: R. Todling
```

```
NAME

 update_ens.csh  - saves updated ensemble member as
                   it completes running

SYNOPSIS

 update_ens.csh  expid memtag type workdir rcfile ncsuffix

where
  expid    -  usual experiment name, e.g., u000\_c72
  memtag   -  current ensemble member
  type     -  file type, e.g. ana or bkg
  workdir  -  work directory where run is taking place
  rcfile   -  typically HISTORY file, otherwise NULL
  ncsuffix -  filename suffix (nc4 or hdf)

DESCRIPTION

 This procedure is responsible for moving the analysis
 and background to their proper location within ENSWORK,
 after they have been generated by the underlying ensemble
 analysis system and the multiple calls the ensemble of GCMs.

 These files will end-up under a directory named updated_ens
 under ENSWORK.

 Example of valid command line:

 update_ens.csh u000_c72 4 bkg \
               /discover/nobackup/rtodling/enswork.1234 nc4

REMARKS

 1. Changing the location within ENSWORK where updated
    members end up is not an option the user can control.
    This is part of the internal mechanisms of the ensemble
    scripts.

AUTHOR
  Ricardo Todling (Ricardo.Todling@nasa.gov), NASA/GMAO
    Last modified: 08Apr2013      by: R. Todling
```

This page intentionally left blank.

# Appendix B

# Acronyms

**3D-Var** three-dimensional variational

**4D-EnVar** four-dimensional ensemble-variational

**ADAS** atmospheric data assimilation system

**AGCM** atmospheric general circulation model

**AOD** Aerosol Optical Depth

**CRTM** Community Radiative Transfer Model

**CVS** Concurrent Versions System

**DFS** degrees of freedom for signal

**DTC** Developmental Destbed Center

**ECMWF** European Center for Medium-Range Weather Forecasts

**EDA** Ensemble of Data Assimilations

**EVADAS** Ensemble-Variational Atmospheric Data Assimilation System

**FP** Forward Processing

**EnKF** Ensemble Kalman Filter

**EnSRF** Ensemble Square-Root Filter

**ESMF** Earth System Modeling Framework

**FGAT** First-Guess at the Appropriate Time

**FP** Forward Processing

**GAAS** Goddard Aerosol Analysis System

**GEOS** Goddard Earth Observing System

**GOCART** Goddard Chemistry Aerosol Radiation and Transport

**GMAO** Global Modeling and Assimilation Office

**GOCART** Goddard Chemistry Aerosol Radiation and Transport

**GSI** Gridpoint Statistical Interpolation

**IAU** Incremental Analysis Update

**LDE** Local Displacement Ensemble

**MERRA** Modern-Era Retrospective Analysis For Research and Applications

**NCEP** National Centers for Environmental Prediction

**NMC** National Meteorological Center

**OBC** observation bias correction

**OMA** observation-minus-analysis

**OMB** observation-minus-background

**PSAS** Physical-space Statistical Analysis System

**TLNMC** Tangent Linear Normal Model Constraint

# Previous Volumes in This Series

**Volume 1** *September 1994*
Documentation of the Goddard Earth Observing System (GEOS) general circulation model - Version 1
**L. L. Takacs, A. Molod, and T. Wang**

**Volume 2** *October 1994*
Direct solution of the implicit formulation of fourth order horizontal diffusion for grid-point models on the sphere
**Y. Li, S. Moorthi, and J. R. Bates**

**Volume 3** *December 1994*
An efficient thermal infrared radiation parameterization for use in general circulation models
**M.-D. Chou and M. J. Suarez**

**Volume 4** *January 1995*
Documentation of the Goddard Earth Observing System (GEOS) Data Assimilation System - Version 1
**James Pfaendtner, Stephen Bloom, David Lamich, Michael Seablom, Meta Sienkiewicz, James Stobie, and Arlindo da Silva**

**Volume 5** *April 1995*
Documentation of the Aries-GEOS dynamical core: Version 2
**Max J. Suarez and Lawrence L. Takacs**

**Volume 6** *April 1995*
A Multiyear Assimilation with the GEOS-1 System: Overview and Results
**Siegfried Schubert, Chung-Kyu Park, Chung-Yu Wu, Wayne Higgins, Yelena Kondratyeva, Andrea Molod, Lawrence Takacs, Michael Seablom, and Richard Rood**

**Volume 7** *September 1995*
Proceedings of the Workshop on the GEOS-1 Five-Year Assimilation
**Siegfried D. Schubert and Richard B. Rood**

**Volume 8** *March 1996*
Documentation of the Tangent Linear Model and Its Adjoint of the Adiabatic Version of the NASA GEOS-1 C-Grid GCM: Version 5.2
**Weiyu Yang and I. Michael Navon**

**Volume 9** *March 1996*
Energy and Water Balance Calculations in the Mosaic LSM
**Randal D. Koster and Max J. Suarez**

**Volume 10** *April 1996*
Dynamical Aspects of Climate Simulations Using the GEOS General Circulation Model
**Lawrence L. Takacs and Max J. Suarez**

**Volume 11** *May 1997*
Documentation of the Tangent Linear and its Adjoint Models of the Relaxed Arakawa-Schubert Moisture Parameterization Package of the NASA GEOS-1 GCM (Version 5.2)
**Weiyu Yang, I. Michael Navon, and Ricardo Todling**

**Volume 12** *August 1997*
Comparison of Satellite Global Rainfall Algorithms
**Alfred T. C. Chang and Long S. Chiu**

**Volume 13** *December 1997*
Interannual Variability and Potential Predictability in Reanalysis Products
**Wie Ming and Siegfried D. Schubert**

**Volume 14** *August 1998*
A Comparison of GEOS Assimilated Data with FIFE Observations
**Michael G. Bosilovich and Siegfried D. Schubert**

**Volume 15** *June 1999*
A Solar Radiation Parameterization for Atmospheric Studies
**Ming-Dah Chou and Max J. Suarez**

**Volume 16** *November 1999*
Filtering Techniques on a Stretched Grid General Circulation Model
**Lawrence Takacs, William Sawyer, Max J. Suarez, and Michael S. Fox-Rabinowitz**

**Volume 17** *July 2000*
Atlas of Seasonal Means Simulated by the NSIPP-1 Atmospheric GCM
**Julio T. Bacmeister, Philip J. Pegion, Siegfried D. Schubert, and Max J. Suarez**

**Volume 18** *December 2000*
An Assessment of the Predictability of Northern Winter Seasonal Means with the NSIPP1 AGCM
**Philip J. Pegion, Siegfried D. Schubert, and Max J. Suarez**

**Volume 19** *July 2001*
A Thermal Infrared Radiation Parameterization for Atmospheric Studies
**Ming-Dah Chou, Max J. Suarez, Xin-Zhong, and Michael M.-H. Yan**

**Volume 20** *August 2001*
The Climate of the FVCCM-3 Model
**Yehui Chang, Siegfried D. Schubert, Shian-Jiann Lin, Sharon Nebuda, and Bo-Wen Shen**

**Volume 21** *September 2001*
Design and Implementation of a Parallel Multivariate Ensemble Kalman Filter for the Poseidon Ocean General Circulation Model
**Christian L. Keppenne and Michele M. Rienecker**

**Volume 22** *August 2002*
Coupled Ocean-Atmosphere Radiative Model for Global Ocean Biogeochemical Models
**Watson W. Gregg**

**Volume 23** *November 2002*
Prospects for Improved Forecasts of Weather and Short-term Climate Variability on Subseasonal (2-Week to 2-Month) Time Scales
**Siegfried D. Schubert, Randall Dole, Huang van den Dool, Max J. Suarez, and Duane Waliser**

**Volume 24** *July 2003*
Temperature Data Assimilation with Salinity Corrections: Validation for the NSIPP Ocean Data Assimilation System in the Tropical Pacific Ocean, 1993–1998
**Alberto Troccoli, Michele M. Rienecker, Christian L. Keppenne, and Gregory C. Johnson**

**Volume 25** *December 2003*
Modeling, Simulation, and Forecasting of Subseasonal Variability
**Duane Waliser, Siegfried D. Schubert, Arun Kumar, Klaus Weickmann, and Randall Dole**

**Volume 26** *April 2005*
Documentation and Validation of the Goddard Earth Observing System (GEOS) Data Assimilation System - Version 4
**Senior Authors: S. Bloom, A. da Silva and D. Dee**
**Contributing Authors: M. Bosilovich, J-D. Chern, S. Pawson, S. Schubert, M. Sienkiewicz, I. Stajner, W-W. Tan, and M-L. Wu**

**Volume 27** *December 2008*
The GEOS-5 Data Assimilation System - Documentation of Versions 5.0.1, 5.1.0, and 5.2.0
**M. M. Rienecker, M. J. Suarez, R. Todling, J. Bacmeister, L. Takacs, H.-C. Liu, W. Gu, M. Sienkiewicz, R. D. Koster, R. Gelaro, I. Stajner, and J. E. Nielsen**

**Volume 28** *April 2012*
The GEOS-5 Atmospheric General Circulation Model: Mean Climate and Development from MERRA to Fortuna
**Andrea Molod, Lawrence Takacs, Max Suarez, Julio Bacmeister, In-Sun Song, and Andrew Eichmann**

**Volume 29** *May 2012*
Atmospheric Reanalyses Recent Progress and Prospects for the Future.
A Report from a Technical Workshop, April 2010
**Michele M. Rienecker, Dick Dee, Jack Woollen, Gilbert P. Compo, Kazutoshi Onogi, Ron Gelaro, Michael G. Bosilovich, Arlindo da Silva, Steven Pawson, Siegfried Schubert, Max Suarez, Dale Barker, Hirotaka Kamahori, Robert Kistler, and Suranjana Saha**

**Volume 30** *September 2012*
The GEOS-ODAS, Description and Evaluation
**Guillaume Vernieres, Michele M. Rienecker, Robin Kovach and Christian L. Keppenne**

**Volume 31** *March 2013*
Global Surface Ocean Carbon Estimates in a Model Forced by MERRA
**Watson W. Gregg, Nancy W. Casey, and Cécile S. Rousseaux**

**Volume 32** *March 2014*
Estimates of AOD Trends (2002–2012) over the World's Major Cities based on the MERRA Aerosol Reanalysis
**Simon Provençal, Pavel Kishcha, Emily Elhacham, Arlindo M. da Silva and Pinhas Alpert**

**Volume 33** *August 2014*
The Effects of Chlorophyll Assimilation on Carbon Fluxes in a Global Biogeochemical Model
**Cécile S. Rousseaux and Watson W. Gregg**

**Volume 34** *September 2014*
Background Error Covariance Estimation using Information from a Single Model Trajectory with Application to Ocean Data Assimilation into the GEOS-5 Coupled Model
**Christian L. Keppenne, Michele M. Rienecker, Robin M. Kovach, and Guillaume Vernieres**

**Volume 35** *December 2014*
Observation-Corrected Precipitation Estimates in GEOS-5
**Rolf H. Reichle and Qing Liu**

**Volume 36** *March 2015*
Evaluation of the 7-km GEOS-5 Nature Run
**Ronald Gelaro, William M. Putman, Steven Pawson, Clara Draper, Andrea Molod, Peter M. Norris, Lesley Ott, Nikki Privé, Oreste Reale, Deepthi Achuthavarier, Michael Bosilovich, Virginie Buchard, Winston Chao, Lawrence Coy, Richard Cullather, Arlindo da Silva, Anton Darmenov, Ronald M. Errico, Marangelly Fuentes, Min-Jeong Kim, Randal Koster, Will McCarty, Jyothi Nattala, Gary Partyka, Siegfried Schubert, Guillaume Vernieres, Yuri Vikhliaev, and Krzysztof Wargan**

**Volume 37** *March 2015*
Maintaining Atmospheric Mass and Water Balance Within Reanalysis
**Lawrence L. Takacs, Max Suarez, and Ricardo Todling**

**Volume 38** *September 2015*
The Quick Fire Emissions Dataset (QFED): Documentation of versions 2.1, 2.2 and 2.4
**Anton Darmenov and Arlindo da Silva**

**Volume 39** *September 2015*
Land Boundary Conditions for the Goddard Earth Observing System Model Version 5 (GEOS-5) Climate Modeling System - Recent Updates and Data File Descriptions
**Sarith Mahanama, Randal Koster, Gregory Walker, Lawrence Takacs, Rolf Reichle, Gabrielle De Lannoy, Qing Liu, Bin Zhao, and Max Suarez**

**Volume 40** *October 2015*
Soil Moisture Active Passive (SMAP) Project Assessment Report for the Beta-Release L4_SM Data Product
**Rolf H. Reichle, Gabrielle J. M. De Lannoy, Qing Liu, Andreas Colliander, Austin Conaty, Thomas Jackson, John Kimball, and Randal D. Koster**

**Volume 41** *October 2015*
GDIS Workshop Report
**Schubert, Siegfried, Will Pozzi, Kingtse Mo, Eric Wood, Kerstin Stahl, Mike Hayes, Juergen Vogt, Sonia Seneviratne, Ron Stewart, Roger Pulwarty, and Robert Stefanski**

**Volume 42** *November 2015*
Soil Moisture Active Passive (SMAP) Project Calibration and Validation for the L4_C Beta-Release Data Product
**John Kimball, Lucas Jones, Joseph Glassy, E. Natasha Stavros, Nima Madani, Rolf Reichle, Thomas Jackson, and Andreas Colliander**

**Volume 43** *September 2015*
MERRA-2: Initial Evaluation of the Climate
**Michael G. Bosilovich, Santha Akella, Lawrence Coy, Richard Cullather, Clara Draper, Ronald Gelaro, Robin Kovach, Qing Liu, Andrea Molod, Peter Norris, Krzysztof Wargan, Winston Chao, Rolf Reichle, Lawrence Takacs, Yury Vikhliaev, Steve Bloom, Allison Collow, Stacey Firth, Gordon Labow, Gary Partyka, Steven Pawson, Oreste Reale, Siegfried D. Schubert, and Max Suarez**

**Volume 44** *February 2016*
Estimation of the Ocean Skin Temperature using the NASA GEOS Atmospheric Data Assimilation System
**Santha Akella, Ricardo Todling and Max Suarez**

**Volume 45** *December 2016*
The MERRA-2 Aerosol Assimilation. NASA Technical Report Series on Global Modeling and Data Assimilation
**C. A. Randles, A. M. da Silva, V. Buchard, A. Darmenov, P. R. Colarco, V. Aquila, H. Bian, E. P. Nowottnick, X. Pan, A. Smirnov, H. Yu, and R. Govindaraju**

**Volume 46** *October 2016*
MERRA-2 Input Observations: Summary and Assessment
**Will McCarty, Lawrence Coy, Ronald Gelaro, Albert Huang, Dagmar Merkova, Edmond B. Smith, Meta Seinkiewicz, and Krzysztof Wargan**

**Volume 47** *May 2017*
An Evaluation of Teleconnections Over the United States in an Ensemble of AMIP Simulations with the MERRA-2 Configuration of the GEOS Atmospheric Model
**Allison B. Marquardt Collow, Sarith P. Mahanama, Michael G. Bosilovich, Randal D. Koster, and Siegfried D. Schubert**

**Volume 48**
*July 2017*

Description of the GMAO OSSE for Weather Analysis Software Package: Version 3
**Ronald M Errico, Nikki C. Privé, David Carvalho, Meta Sienkiewicz, Amal El Akkraoui, Jing Guo, Ricardo Todling, Will McCarty, William M. Putman, Arlindo da Silva, Ronald Gelaro, Isaac Moradi**

**Volume 49**
*March 2018*

Preliminary evaluation of influence of aerosols on the simulation of brightness temperature in the NASA Goddard Earth Observing System Atmospheric Data Assimilation System
**Jong Kim, Santha Akella, Arlindo M. da Silva, Ricardo Todling, Will McCarty**