One approach for using offline DA to generate ocean initial condition perturbations

Dan Amrhein

October 15, 2020

Our goal is to generate seasonally varying perturbations $\mathbf{p}^j$, where $j = \{1, 2, \ldots, 12\}$ indexes different monthly values, that reflect the uncertainty of FOSI in representing the true ocean state at the time of prediction initialization. Unlike approaches (e.g., using singular vetors) targeting unstable modes to yield rapid *growth* in ensemble spread, the goal is to have an accurate *initial* spread reflecting the uncertainty of FOSI ICs about true ocean conditions, similar to what would be provided by an ensemble DA posterior.

Here we generate $\mathbf{p}^j$ based on estimated statistics of misfits between FOSI output and observations (or observational products). Let $\mathbf{x}_{\text{FOSI}}(t_k)$ be the ocean state vector in the FOSI simulation at time t_k, and $\mathbf{y}(t_k)$ the set of observations available at that time. The misfit in observation space is given by

$$\mathbf{m}(t_k) = \mathcal{H}(t_k)\, \mathbf{x}_{\text{FOSI}}(t_k) - \mathbf{y}(t_k), \tag{1}$$

where $\mathcal{H}$ is the observation forward operator, which can vary in time as observational availability changes. One might compute the statistics of misfits based on multiple realizations of $\mathbf{m}(t_k)$ and use this to generate observation-space perturbations. However, Gauss-Markov interpolation of misfits in space allows us to 1) provide complete spatial coverage for perturbations and 2) project them into the space of model variability (with the goal of reducing initialization shock from off-attractor features). This is equivalent to an offline computation of the Kalman increment $\tilde{\mathbf{x}}(t_k)$ at each time step of the FOSI run,

$$\tilde{\mathbf{x}}(t_k) = \mathbf{K}\mathbf{m}(t_k) \tag{2}$$

where the gain is given by

$$\mathbf{K} = \mathbf{B}^j \mathcal{H}^\top \left[\mathcal{H}\mathbf{B}^j\mathcal{H}^\top + \mathbf{R}\right]^{-1} \tag{3}$$

and $\mathbf{B}^j$ is an estimate of the prior state covariance for the month (indexed by the j^{th} month in which t_k falls). This covariance could be estimated, e.g. for January, using the EnOI procedure of taking multiple days from the FOSI run in January. We can then compute increments over different N time intervals within the j^{th} month, and over different years of FOSI output, yielding a set of $\tilde{\mathbf{x}}^j(t_k)$ that we can concatenate into a matrix $\tilde{\mathbf{X}}^j$,

$$\tilde{\mathbf{X}}^j = \left[\tilde{\mathbf{x}}^j(t_1), \tilde{\mathbf{x}}^j(t_2), \ldots, \tilde{\mathbf{x}}^j(t_N)\right]. \tag{4}$$

To generate perturbations $\mathbf{p}^j$ having the same spatial statistics as $\tilde{\mathbf{x}}^j$, compute the singular value decomposition of $\tilde{\mathbf{X}}^j$,

$$\mathbf{U}^j \mathbf{S}^j \left(\mathbf{V}^j\right)^\top = \tilde{\mathbf{X}}^j. \tag{5}$$

and define $\mathbf{w}$ to be a random vector of length N (corresponding to the rank of $\tilde{\mathbf{X}}^j$) drawn from the joint standard normal distribution whose covariance is the identity matrix. A perturbation can be generated as

$$\mathbf{p}^i = \mathbf{U}^j \mathbf{S}^j \mathbf{w} \tag{6}$$

and additional realizations of $\mathbf{w}$ used to generate a set of i.i.d. perturbations used to generate initial ensemble spread.

Questions:

- What time interval should be used for estimated increments?
- We can expect that the ensemble of increments used to generate perturbation statistics will have a nonzero ensemble mean. Should this be removed? One could explore adding it back in to final forecasts?
- What time intervals should be used to compute statistics? Do we run into trouble using statistics of future variability to initialize hindcast experiments?
- Which observations and/or observational products to use? What should we use for the observational covariance $\mathbf{R}$?