Development of a hybrid variational-ensemble data assimilation technique for observed lightning tested in a mesoscale model

Karina Apodaca[1], Milija Zupanski[1], Mark DeMaria[2,*], John A. Knaff[2], and Lewis D. Grasso[1]

[1]{Colorado State University/Cooperative Institute for Research in the Atmosphere, Fort Collins, Colorado, USA}
[2]{NOAA Center for Satellite Research and Applications, Fort Collins, Colorado, USA}
[*]{now at: Technology and Science Branch, National Hurricane Center, Miami, Florida, USA}
Correspondence to: K. Apodaca (karina.apodaca@colostate.edu)

Abstract

Lightning measurements from the Geostationary Lightning Mapper (GLM) that will be aboard the Goestationary Operational Environmental Satellite – R Series will bring new information that can have the potential for improving the initialization of numerical weather prediction models by assisting in the detection of clouds and convection through data assimilation. In this study we focus on investigating the utility of lightning observations in mesoscale and regional applications suitable for current operational environments, in which convection cannot be explicitly resolved. Therefore, we examine the impact of lightning observations on storm environment. Preliminary steps in developing a lightning data assimilation capability suitable for mesoscale modeling are presented in this paper. World Wide Lightning Location Network (WWLLN) data was utilized as a proxy for GLM measurements and was assimilated with the Maximum Likelihood Ensemble Filter, interfaced with the Nonhydrostatic Mesoscale Model core of the Weather Research and Forecasting system (WRF-NMM). In order to test this methodology, regional data assimilation experiments were conducted. Results indicate that lightning data assimilation had a positive impact on the following: information content, influencing several dynamical variables in the model (e.g. moisture, temperature, and winds), improving initial conditions during several

data assimilation cycles. However, the 6 h forecast after assimilation, did not show a clear improvement in terms of root mean square errors.

1 Introduction

Thunderstorms are an important component of the climate system as they can impact the atmospheric environment around them; they are capable of redistributing moisture, heat, and wind patterns (Price, 2013). The assimilation of lighting observations is a relatively new field. Several efforts to incorporate lightning data into Numerical Weather Prediction (NWP) models have been made recently (Alexander et al., 1999, Papadopoulos et al., 2005, Mansell et al., 2007, Pessi and Bussinger, 2009, Fierro et al., 2012). In the vast majority of these studies dynamical relaxation, or nudging techniques were applied. Even though these studies highlighted the importance of utilizing lightning observations to improve the representation of convection in models, they had less emphasis on improving the environmental conditions.

Motivated by the initial success of nudging techniques in cloud-resolving model applications, the objective of this study is to investigate if lightning observations can be useful in mesoscale, regional, and global applications at a coarse resolution, in which convection cannot be explicitly resolved. Therefore, we would like to evaluate the impact of lightning observations on the environment around storms, with potential implications to data assimilation, reanalysis, and climate studies. As for any other observation, the information from lightning observations can have impacts at several spatiotemporal scales. In the case of lightning, one can assume that most of the information relates to cloud-resolving processes. However, there should be also a fraction of lightning information that can spread into larger scales (e.g., the storm environment). In this study we will evaluate the large-scale component of information from lightning observations.

We anticipate that a myriad of applications can stem from monitoring lightning activity. For instance, the lack of ground-based observations (e.g. radiosondes, radars, etc.) over the open oceans can result in deficient initialization of numerical weather and climate prediction models, especially if weather systems that develop in these regions subsequently travel to continental landmasses. Satellite radiances are an important source of observations over the oceans. However, processing satellite observations requires considerably more computational time due to the use of radiative transfer models, rather than just processing lightning observations, which is computationally less intensive. Therefore, the incorporation of this new type of data can provide useful information for model initialization.

In addition, lightning may have a significant impact on the Earth's climate by producing nitrogen oxides (NO_x) in the upper troposphere. NO_x is a precursor of ozone, a major green house gas and pollutant (Price, 2013, Barthe et al., 2010). The predicted concentrations of lightning-NO_x from NWP models coupled with chemistry still contain large uncertainties. Incorporating geo-located lightning data may assist these models in the simulation of convection, and consequently NO_x production.

Lightning might be useful in future climate change monitoring studies due to the interplay between lightning and atmospheric parameters, such as, temperature, upper tropospheric water vapor, and cloud cover (Price, 2013). Since lightning can be easily monitored through surface networks and satellite platforms it can be a useful tool for tracking changes in important climate parameters in the future (Price, 2009).

Satellite instruments have been launched in the past with the objective of studying storm dynamics, cloud characteristics, annual and inter-annual variability of thunderstorms, etc. (Adamo et al., 2009). In 1997, the Lightning Imaging Sensor (LIS) was launched aboard de joint National Aeronautics and Space Administration (NASA) and the Japan Aerospace Exploration Agency (JAXA) Tropical Rainfall Measuring Mission (TRMM). This instrument can detect lightning activity continuously at a horizontal resolution of 4 km over the tropics (http://trmm.gsfc.nasa.gov/overview_dir/lis.html).

In the near future, mapping of lightning from geostationary orbit at cloud scale resolution will be possible, thus complementing established surface detection networks (Adamo et al., 2009, Finke, 2009). The launch of the

Geostationary Lightning Mapper (GLM) instrument that will be aboard the next generation of the National Oceanic and Atmospheric Administration (NOAA) geostationary satellites (i.e., GOES–R, http://www.goes-r.gov/spacesegment/glm.html) will allow continuous day and night monitoring of total lightning activity over the Americas and adjacent ocean regions up to 52 degrees north. One of the advantages over previous lightning mapping instruments is that it will be able to monitor weather affecting the adjacent ocean regions of the continental United States and not just the tropics. Some of the mission objectives for the GLM instrument include: improvement in severe thunderstorm lead times and false alarm reduction, advancements in the initialization of NWP models through better identification of deep convection, creation of lightning climatologies to track decadal changes in lightning activity, among others (Adamo et al., 2008).

In this paper the possibility of assimilating lightning observations within a hybrid variational-ensemble system in a mesoscale numerical weather prediction model is explored, focusing on the typical resolution of operational weather forecasting and climate models. The methodologies presented herein represent an initial stage towards developing a comprehensive, multivariate, multi-scale, multi-sensor data assimilation system that prepares for the assimilation of lightning data along with other types of observations.

Eventually, this data assimilation technique will be tested in different applications at various time and length scales. In the mean time, we intend to investigate if the assimilation of lightning data can (1) add information content into a mesoscale modeling system that can resolve a convective environment, rather than explicit convection, (2) positively impact the dynamical variables of the model, and (3) improve analysis and prediction. Note that a coarse resolution is also typical of climate models, and thus assessing the utility of lightning observations in data assimilation at these scales can be relevant for climate studies as well. To our knowledge, lightning data have not been used in operational weather prediction, in climate monitoring studies, or in reanalysis. By assimilating lightning data in a coarse resolution model we are taking first steps toward extending their use to weather and climate applications.

As a proof of concept case we chose the mesoscale convective system that spawned numerous tornados over the southeastern United States on 27-28 April 2011. Lightning data from the World Wide Lightning Location Network (WWLLN, http://webflash.ess.washington.edu) was used as a proxy to test the potential impact of the assimilation of lightning flash rates measured by the GLM. This data network has global coverage, including ocean regions. For North America, this lightning detection network better approximates the coverage of the upcoming GLM instrument compared to some surface networks that primarily cover the continental United States.

The data assimilation system (DA) used in this study was the Maximum Likelihood Ensemble Filter (MLEF – Zupanski, 2005; Zupanski et al., 2008), which was interfaced with the non-hydrostatic core of the Weather and Research Forecasting system (WRF-NMM - Janjić et al., 2010). The simplified microphysics and low-resolution of the model defined the spatiotemporal scales for data assimilation, as well as the options for the employed observation operator. In this case, a 6-hour data assimilation window was chosen (±3 hours from a central time), in which the lightning observations were averaged at a horizontal resolution of 10 km closely matching that of the innermost domain of WRF-NMM.

This paper is organized in the following manner: the methodology for using lightning observations is described in Sect. 2, details on the experimental design are provided in Sect. 3, followed by results in Sect. 4, and finally a summary and future work are presented in Sect. 5.

2 Methodology for utilizing lightning observations

2.1 Data Assimilation System

WRF-NMM was interfaced with MLEF, a hybrid ensemble-variational data assimilation method developed at Colorado State University. The solution of the analysis maximizes the likelihood of the posterior probability distribution, obtained by a minimization of a cost function that includes a general nonlinear observation operator. As in typical variational and ensemble data assimilation methods, a cost function is derived using a Gaussian probability density function framework. Like other ensemble data assimilation algorithms, MLEF produces an estimate of the analysis uncertainty (e.g., analysis error covariance). In addition to the common use of ensembles in calculations of the forecast error covariance, the ensembles in MLEF are exploited to efficiently calculate the Hessian preconditioning and the gradient of a cost function. The MLEF method is well suited for use with highly nonlinear observation operators, for a small additional computational cost of the minimization procedure. Relevant prognostic WRF-NMM variables were selected as control variables, as they can significantly impact the initial conditions, which can, in turn, influence the forecast. This selection includes the following variables: temperature (*T*), specific humidity (*q*), hydrostatic pressure depth (*PD*), the *U* and *V* components of the wind, and Cloud Water Mass (CWM – total cloud condensate in WRF-NMM) that combines all cloud hydrometeors into a total sum. The goal is to minimize the following cost function:

$$J(x)=\frac{1}{2}\left[x-x^f\right]^T \mathbf{P}_f^{-1}\left[x-x^f\right]+\frac{1}{2}\left[y-h(x)\right]^T \mathbf{R}^{-1}\left[y-h(x)\right] \quad (1)$$

where x represents the above defined control variables with a forecast error covariance $\mathbf{P}_f$, the index f denotes the forecast guess, y is the lightning flash rate observations with an error covariance $\mathbf{R}$, and h is the nonlinear lightning observation operator that maps the control variables to the lightning flash rate observations. The superscript T indicates the transpose of a matrix.

2.2 Lightning flash rate observations

Since the actual lightning measurements are lightning strikes, while the lightning observation operator is commonly related to lightning flash rates, it was necessary to transform lightning strikes into flash rates. In doing so, a subset domain containing all lightning strikes was defined and subsequently partitioned into a rectangular horizontal grid (different from the model grid), with a spacing of 0.1 degrees (~10 km) in order to be comparable with the horizontal grid spacing of the smallest domain of our model configuration that will be discussed in Sect. 3.2. The choice of a regular grid that is not identical to the model grid is arbitrary. In our case, it was motivated by a desire to keep the observation information formally independent from the model, i.e. to not use any information about the model when defining observations and observation errors. Lightning strikes counted in each local area surrounding a grid point during a 6-hour time window coinciding with the data assimilation interval were assigned to a particular grid point, and then divided by a time interval to form lightning flash rates. Therefore, the lightning flash rate observations are grid-point values that represent a cumulative count of geo-located lightning strikes over the 6-hour assimilation time window (±3 hours from a central time), rather than the instantaneous measurements. Note that the observed lightning flash rates were assumed to be greater than zero, i.e., the observation grid points without any lightning strikes were not included in the observations pool. Observations of zero lightning can be important in pointing the location of misplaced convection events. However, it is not clear how this information would impact convection events that are not characterized by strong lightning. It is likely that additional information would be needed in order to selectively define zero lightning observations. Even though, this information is important, it needs further investigation. The non-negative character of lightning observations introduces a skewness that points out to a need for a non-Gaussian PDF in lightning data assimilation (e.g., Fletcher and Zupanski, 2007; Lien et al. 2013). This issue will be examined in the future since it can potentially improve the utility of lightning data.

2.3 Lightning flash rate observation operator

The lightning flash rate observation operator h (Eq. (1)) includes two operations: a transformation (h_2) and an interpolation (h_1), i.e. $h = h_1h_2$. In this study the forward lightning transformation operator (h_2) was adopted by exploiting the relationship between lightning and vertical velocity. This choice was influenced by the properties of a bulk microphysics scheme used in the WRF-NMM model (e.g., Ferrier, 2005), and by the coarse assimilation time window that effectively restricts using the cloud-scale information about hydrometeors and their interactions. A bi-linear interpolation technique was used to interpolate the guess lightning flash rates to observation location (h_1).

As seen in previous studies, lightning is related to updrafts that support a deep layer of super-cooled water droplets and a mixed phase region where charge separation occurs (Black and Hallet, 1999). Based on Price and Rind (1992), an empirical relationship between maximum updraft velocity ($w_{\max}$) and lightning flash rate (f) given by:

$$h_2 = f = cw_{\max}^{\beta} \tag{2}$$

was used, under the assumption that updrafts are positively correlated to cloud top height. $c = 5 \times 10^{-6}$ and $\beta = 4.5$ are empirical parameters. β is a value derived from satellite data climatologies for continental clouds as in Price and Rind (1992). Both c and β are dimensionless.

The procedure to develop the lightning observation operator started with an approximate calculation of vertical velocity from WRF-NMM, through the use of a reduced version from the nonhydrostatic continuity equation

$$w \approx \frac{1}{g}\left(\frac{\partial\Phi}{\partial t} + v \bullet \nabla_{\sigma}\Phi + \dot{\sigma}\frac{\partial\Phi}{\partial t}\right) \tag{3}$$

where w is the vertical velocity, g is the gravity constant, Φ is the geopotential, $\boldsymbol{v}$ is the horizontal wind vector, and $\dot{\sigma}$ is the vertical velocity in a sigma coordinate (Janjić, 2005). An approximation was required because vertical velocity is not a predictive, but rather a diagnostic variable in WRF-NMM. After an approximate value of vertical velocity was obtained, the maximum vertical velocity was calculated for horizontal points according to the following procedure: values of Cloud Water Mass (CWM - total cloud condensate in WRF-NMM) CWM ≥ 10^{-5} (kg kg^{-1}) were searched for at each model grid point and surrounding neighboring points along all vertical model levels. We defined a 5 x 10 grid point area (approximately a square domain in Arakawa E-grid staggering used in WRF-NMM) surrounding the central point in order to introduce a smooth transition for the calculation of w_{max}. This procedure was applied to avoid taking into account values of w_{max} in regions without clouds. If the CWM threshold was reached, the value of w_{max} was calculated at a grid point and surrounding points at all vertical levels, otherwise w_{max} was set to zero. Once the value of w_{max} was calculated, it was possible to calculate values of lightning flash rate from Eq. (2). Since the calculation of w (e.g., Eq. (3)) and w_{max} includes prognostic model variables, all control variables can impact lightning flash rates.

Since both a new observation type (lightning flash rate) and an untested observation operator (Eq. (2)) were introduced into the data assimilation system, statistics of innovation vectors (observation minus guess) of lightning flash rates needed to be examined first. Figure 1 shows the statistics of the normalized innovation vectors $R^{-1/2}\left[y-h(x^f)\right]$ at several observation times. A skewed histogram of the Probability Distribution Function (PDF) innovation vectors (left) can be readily seen, implicitly indicating that the observed values of lightning flash rate were considerably larger than the guess. Therefore, it was necessary to perform a correction. An option could have been to increase the value of parameter c in Eq. (2) to reduce the skewness. However, trial experiments indicated a large uncertainty of the parameter c from one observation time to another, in occasions ranging over two orders of magnitude. In order to deal with this error of the observation operator (Eq. (2)), an adjustable multiplicative correction parameter (α >0) was included so that h_2 would become αh_2. At each observation time an optimal parameter α_{opt} was estimated by minimizing the following cost function:

$$J(\alpha)=\frac{1}{2}\left[\log(\alpha)-\log(\alpha_0)\right]^T \mathbf{W}^{-1}\left[\log(\alpha)-\log(\alpha_0)\right] + \frac{1}{2}\left[\log(y)-\log(\alpha h(x^f))\right]^T \mathbf{R}_L^{-1}\left[\log(y)-\log(\alpha h(x^f))\right] \tag{4}$$

where $\mathbf{R}_L$ is the observation error covariance associated with a logarithmic transformation, α_0 is a guess value, and **W** is the uncertainty matrix of the guess value. The choice of a logarithmic transformation was influenced by the fact that lightning flash rate is strictly positive definite and that such procedure could better deal with the large uncertainty of the parameter α. As shown in the appendix (Sect. 7), the solution of α_{opt}, which minimizes the cost function, i.e., Eq. (4), is given by:

$$\alpha_{opt}=\exp\left[\frac{\frac{1}{N_{obs}}\sum_{i=1}^{N_{obs}}\log\left(\frac{y}{h(x)}\right)_i}{1+\frac{r_0}{w_0}}\right]. \tag{5}$$

where N_{obs} is the number of observations, $diag(\mathbf{W})=w_0$ and $diag(\mathbf{R}_L)=r_0$. Therefore the lightning observation transformation operator (Eq. (2)) was substituted by

$$h_2=f=\alpha_{opt}cw_{\max}^{\beta}\,. \tag{6}$$

The observation operator transformation (e.g., Eq. (6)) is defined over a 2-dimensional horizontal domain only since flash rate f is a horizontal field (e.g., number of hits per area and time). This requires w_{max} to be 2-dimensional as well. Therefore, w_{max} is defined for each horizontal grid point, as the maximum value of vertical velocity (w) over all vertical levels. The flow diagram of the data assimilation system and the lightning observation operator are illustrated in Fig. 2.

2.4 Information content of lightning observations

In general terms, the impact of observations can be quantified using an uncertainty reduction after data assimilation. Since entropy measures the uncertainty, one can use the formalism of Shannon information theory (Shannon and Weaver 1949) to define information content of observations as an entropy difference before and after data assimilation. As shown in Rodgers (2000), the entropy is considerably simplified with a Gaussian probability assumption and information content can be conveniently expressed in terms of degrees of freedom for signal (d_s),

$$d_s = trace\left[\mathbf{I} - \mathbf{P}_a \mathbf{P}_f^{-1}\right], \tag{7}$$

where *trace* is the trace function, **I** is the identity matrix, and $\mathbf{P}_a$ is the analysis error covariance. This can be further reduced in terms of the eigenvalues of the observation information matrix, given by:

$$\mathbf{P}_f^{T/2}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\mathbf{P}_f^{1/2} = \mathbf{U}\Lambda\mathbf{U}^T \tag{8}$$

where Λ and **U** are the eigenvalues and eigenvectors matrices, respectively, and **H** is the Jacobian of the observation operator. The degrees of freedom for signal are then

$$d_s = \sum_i \frac{\lambda_i^2}{1+\lambda_i^2} \tag{9}$$

where λ_i are the eigenvalues. Zupanski et al. (2007) showed that this formula could also be useful in reduced-rank, ensemble space calculations, in which the summation is performed over the number of ensemble members. Since an eigenvalue decomposition of the observation information matrix is a component of the MLEF algorithm, additional cost of calculating d_s is minimal. By calculating the degrees of freedom for signal we can quantify the impact of the lightning observations in terms of an uncertainty reduction. Note that Eq. (9) has non-negative values between 0 and N_{ens}, depending on the structure of the observation information matrix. If there is a negligible impact of lightning observations the number of degrees of freedom for signal will be close to zero, i.e. much smaller than the number of ensemble members.

3 Experimental Design

3.1 General synoptic description of the case study

As a proof of concept case for regional lightning data assimilation over a continental area we selected the severe weather event that occurred on 27-28 April 2011, where an estimated 292 tornadoes hit the southeastern, mid-west and northeast United States, according to the Storm Report Center (Fig. 3, top panel). A figure of 500 hPa heights, with color contours of wind speed and surface observations from the Forecast Systems Laboratory (Fig. 3, bottom panel) shows that atmospheric conditions created a perfect scenario for severe weather development. An upper-level low centered on Minnesota along with the advance of a deep trough and its associated jet streak (wind speed exceeding 41.15 m sec^{-1}) aloft led to rapid atmospheric destabilization in the afternoon of 27 April. Surface moist-warm flow arrived from the Gulf of Mexico, with dew points exceeding 21 °C and wind gusts over 7.72 m sec^{-1} at the Alabama coast. An upper level disturbance sparked a broad area of showers and thunderstorms as it moved across the frontal boundary on the previous evening. The eastern edge of this line of showers and storms continued to move eastward, in concert with the upper-level disturbance, reaching the northwest Alabama border around 0700 UTC on the 27^{th}. Meanwhile, surface winds backed to the south-southeast as the disturbance moved into the area, while winds at the 850 hPa level (around 1,500 m) increased to 26-28 m sec^{-1} and became more southerly. The combination of high low-level moisture and increasing shear provided the setup for damaging winds, large hail and brief tornadoes. This line experienced further intensification as it moved into northwest Alabama, especially after 0900 UTC. This line of severe storms pushed into northwestern Alabama prompting a tornado watch for all of northern Alabama and portions of southern middle Tennessee until 1400 UTC. A deep layer shear and moisture increased dramatically later in the afternoon and evening of the 27^{th} ahead of the strong cold front. This combination of strong instability and high shear continued through the evening hours ahead of the cold front before it pushed east of the area into Georgia. This produced the last and most violent round of severe weather, which began around 2030 UTC for northern Alabama as supercells began to line up to the southwest of the area. During the early afternoon hours, the potential for destructive tornadoes was highlighted by the Storm Prediction Center's upgrade to a rare High Risk for severe weather around 1300 UTC. This prompted a Particularly Dangerous Situation (PDS) tornado watch, which was issued for northern Alabama and portions of southern middle Tennessee at 1945 UTC. The potential really ramped up from noon through 2100 UTC. During this period, much of Alabama experienced numerous supercell thunderstorms producing strong to violent tornadoes, including five EF-4 tornadoes and one EF.5 in the Huntsville Forecast Area (NOAA Service Assessment, Hayes, 2011, http://www.srh.noaa.gov/hun/?n=hunsur2011-04-27_setup).

3.2 Model and domain configuration

The WRF-NMM version 3 model from the Developmental Testbed Center [http://www.dtcenter.org] was employed in this study. WRF-NMM was developed by the NOAA/National Centers for Environmental Prediction (NCEP) (Janjić et al., 2010). For simplicity, only some physics and dynamics choices are mentioned. The microphysics option was Ferrier (Ferrier, 2005), which includes prognostic mixed-phase processes. The longwave and shortwave radiation options were the Geophysical Fluid Dynamics Laboratory (GFDL) schemes. The GFDL longwave radiation scheme includes the transmission and absorption of carbon dioxide, ozone, and water vapor in multiple spectral bands. Likewise, in the GFDL shortwave scheme, ozone and water vapor are the main absorbers. Both schemes include cloud microphysical effects (Falkovich et al., 2005). The planetary boundary layer option was the Mellor-Yamada-Janjinc (Janjić, 1994). The land surface option was the NOAH Land-Surface model (Ek et al., 2003) with soil temperature and moisture in four layers, fractional snow cover and frozen soil physics. For the cumulus parameterization, Betts-Miller-Janjić was selected. This scheme adjusts deep shallow convection with a relaxation towards variable humidity and temperature profiles (BMJ-Janjić 1994, 2000).

The WRF-NMM simulations in this study were configured with two domains. Domain 1 (D01) had a horizontal grid spacing of 27 km and a size of 1350 by 2592 km^2 (50 x 96 grid points). This domain covered parts of the mid-west, the Gulf of Mexico, the Atlantic Ocean, and the eastern United States. Domain 2 (D02), centered on Alabama, had a horizontal grid spacing of 9 km and a size of 540 by 1170 km^2 (60 x 130 grid points) (Fig. 4). Both domains had a vertical extent of 27 vertical levels.

3.3 Data sets and data assimilation system setup

The ensemble boundary conditions are obtained from the NCEP Global Forecast System (GFS) using the WRF preprocessing system (WPS). With the exception of the initial ensemble preparation (i.e. cycle0 in our terminology), the initial conditions for the ensemble members are obtained through the MLEF algorithm by adding the analysis square root error covariance columns to the analysis. Further information about the MLEF methodology can be found in Zupanski (2005) and Zupanski et al. (2008). The localization setting for the ensemble-based covariance includes a de-correlation length of 90 km. The data assimilation period starts at 1800 UTC 26 April 2011, ending on 1200 UTC 28 April 2011. Note that there is no data assimilation at the initial time.

In the present study, WWLLN data were assimilated. The WWLLN is an experimental lightning detection network that provides the location of cloud-to-ground (CG) and some intra-cloud lightning (IC) strikes in real-time, it has a global coverage with 10 km location accuracy and flash detection accuracy greater than 50% (Lay, 2004). WWLLN is for the most part; a time average of geo-located CG lightning flashes that cannot address the cloud-resolving characteristics of lightning. Nonetheless, for the purposes of evaluating the impact of lightning observations on the storm environment, making a distinction between CG and IC lightning is beyond the scope of this study. The ensemble size was set to 32 in order to match the number of processors per node, with a data assimilation interval of 6 hours to match the frequency of the Global Forecast System (GFS) input files. The 6-hourly averaged lightning flash rates (±3 hours) were assimilated at each central time t_n (n>0). An initial 6-hour forecast was obtained at cycle0 from WRF-NMM with the GFS files (from t_{n-3h} to t_{n+3h}) and it was used as a first guess to obtain the analysis solution for the next cycle. The background state x^f, or prior, is an estimate of the most likely dynamical state; it is a deterministic forecast from the previous assimilation cycle. The analysis solution was obtained as a maximum likelihood estimate from the assimilation of observations at the central time t_n (Zupanski, 2005). These steps were repeated during each cycling period. Figure 5 shows the data assimilation timeline. The observational error was assumed to be 0.10 hits km^{-2} h^{-1}.

3.4 Description of the experiments

Three simulations were performed to assess the impact of the assimilation of lightning flash rates into a mesoscale NWP:

1. The first experiment was a single observation test (1-OBS), performed to evaluate the impact of assimilating lightning flash rates at a single WWLLN location (34.5°N, 89°W) on the analysis increment (analysis minus background) of a subset of the control variables (q, T, U, and V) mentioned in Sect. 2.1 and to implicitly illustrate the complex structure of the flow-dependent forecast error covariance. The difference between the initial observation and the guess was assumed to be one standard deviation of the observation error covariance **R**, i.e., $y = x^f + \sigma_R$ where $\sigma_R = 1$.
2. The second experiment was a control run, without the assimilation of lightning data, referred to as no-data-assimilation (NODA). Note, however, that lightning observations were still present in the simulation in order to define the optimal regression parameter α_{opt}.

3. In addition to the two simulations mentioned before, an experiment that included the assimilation of WWLLN lightning data (LIGHT) was performed. LIGHT had the same set-up as the NODA simulation; the only difference was the assimilation of lightning flash rates.

4 Results

In the following sections, we present an evaluation of the impact of the assimilation of lightning data for the 27-28 April 2011 severe weather event focusing on domain D02 (9km resolution). First, results of the (1-OBS) experiment are shown, followed by an evaluation of the time-flow-dependent forecast error covariance through the use of degrees of freedom for signal to quantify the information added to the system by the assimilation of the lightning observations. Then an evaluation of several synoptic fields from the LIGHT simulation and validation of the DA system through comparisons with some observations are presented. Thereafter, an assessment between the LIGHT and NODA simulations through the calculation of Root Mean Square (RMS) errors of the lightning observations is shown

4.1 1-OBS experiment

The difference between the analysis and the 6-hour forecast (background) was evaluated. Figure 6a shows the 700 hPa analysis increments of specific humidity (*q*) at 1800 UTC 27April 2011, or cycle 3 in the data assimilation period. This time was chosen since tornados started developing over northern Alabama just a couple of hours before. The black dot indicates the location of the single observation being assimilated (34.5°N, 89°W). A clear dipole of positive and negative analysis increments in q, with a magnitude of $\pm$ 4 x 10^{-5} kg kg^{-1}, is observed at opposite sides of the location of the single observation. The analysis increment of temperature (*T*) at 700 hPa (Fig. 6b) shows regions of positive and negative analysis increments, with a magnitude of $\pm$ 4 x 10^{-2} degrees K, over the same regions as *q*, but with opposite sign. The plot of wind speed at 700 hPa (Fig. 6c) shows a positive analysis increment of 2.7 x 10^{-1} m sec^{-1} with maximum values coinciding with the region of positive potential temperature increment. The spatial extension of the impact of assimilating a single lightning strike on some of the dynamical variables of the model in D02 (9 km resolution) was: (i) on specific humidity the impact extends to approximately 12 grid points (~ 110 km), (ii) for temperature to 20 grid points (~ 180 km), and (iii) for wind approximately 30 grid points (~ 270 km).

The former Fig. (6a, 6b, and 6c) indicates that the assimilation of lightning at a single location impacted the atmospheric environment at surrounding grid points. The magnitude of the analysis increments indicates non-negligible adjustments on dynamical variables of the mesoscale model. Most importantly, it can be noted that the hybrid DA system was able to spatially spread the information of a single lightning observation and influence the initial conditions of specific humidity, temperature, the *U* and *V* components of the wind and other control variable elements. These results are a manifestation of the complex structure of the ensemble forecast error covariance matrix. This is important since it indicates that the information from lightning observations can impact the initial conditions and eventually the forecast of coarse resolution models.

4.2 Evaluation of information content of the lightning observations

In these experiments, the degrees of freedom for signal were computed in ensemble subspace following Zupanski et al. (2007). The top-three plots in Fig. 7 show degrees of freedom for signal during three assimilation cycles (1, 2 and 3, as an example) and observed GOES-IR and lightning flash rates at matching times (bottom-three plots). The areas of highest density of WWLLN lightning observations are in agreement with information content, implying that the time-flow-dependent forecast error covariance had a direct relationship to the observations throughout the assimilation period. Maximum values of degrees of freedom for signal of 12, 22, and 10 for cycles 1, 3, and 5, respectively can be observed in Fig. 7. These values indicate that the benefit of the observations is important, otherwise these values would be close to zero, i.e. much smaller than the number of ensemble members, 32 in this case. On the other hand, if the former values were to approach the number of ensemble members, this would be an indicator of the introduction of noise to the DA system by the observations and their possible benefit would be nullified. Note however, that the agreement in cycle 3 was not very good. It is possible that ensemble perturbations where not large enough over northeastern Alabama where another maximum was missing. This lack of agreement can arise from the use of a reduced rank ensemble approach and consequently not having enough spread in the ensembles. However, the agreement improved in subsequent cycles (e.g., shown for cycle 5).

4.3 Impacts on the environment during the severe weather event

The following results correspond to 0000 UTC 28 April 2011, cycle 5 in the data assimilation time line, at the time when an EF4 tornado affected Tuscaloosa and Birmingham, Alabama. Fields of wind, absolute vorticity and Convective Available Potential Energy (CAPE) from both experiments (LIGHT and NODA) portray a distinctive scenario of an environment favorable for the strengthening of deep convection, but with some differences. Figure 8a shows background (forecast) winds at 850 hPa for the NODA experiment. Figure 8b shows background winds at 850 hPa for the LIGHT experiment. A core of increased wind speed over northern Alabama can be observed in both plots. However, the core of maximum wind speed has a larger spatial coverage in the LIGHT experiment and based on computed differences, stronger winds with a magnitude, in the order of 4 to 6 m sec^{-1} were found in the LIGHT experiment. Note that this region is co-located with an area of high density of WWLLN lightning observations (Fig. 8c). Figures 9a and 9b correspond to the analysis increment of the 850 hPa winds and absolute vorticity, respectively. Regions of positive increments are found near the left-hand side in both plots (4 to 6 m sec^{-1}

in wind speed and 4 x 10^{-4} sec^{-1} in vorticity). Almost no analysis increments can be found in the region where the densest lightning observations are located (Alabama). Among possible reasons, we can mention the following: (*i*) the largest forecast uncertainty (i.e. ensemble perturbations) typically occurs in the areas of strongest dynamical instability, in this case, in the region where a dry line was present over the states of Louisiana, Mississippi, Arkansas, and Missouri. Even though, the dry line may not be characterized by the strongest lightning activity, there were still some isolated lightning observations present over the domain as seen in Fig. 8c, (*ii*) alternatively, it may be a consequence of using an ensemble-based forecast error covariance that was not able to produce sufficient uncertainty in all relevant areas.

Similarly, by analyzing CAPE at the forecast step for both experiments (Fig. 10a,b), a region of high CAPE gradient is observed on the left hand side of the domain, indicating the presence of a well-defined dry line. However, no significant differences were found between both experiments for this particular assimilation cycle (cycle 5). One possible reason is that there were no lightning observations present at the core where the strongest CAPE was observed. Therefore, lightning was not able to significantly impact CAPE. Further investigation is required to see if the same behavior occurs for other cycles and case studies.

Forecast CAPE was validated by comparing the model output with observations from the Storm Prediction Center's Surface Mesoanalysis at 40 km resolution. Figure 10c shows observed CAPE. A well-defined dry-line can be readily seen in the plot of background CAPE (Fig. 10a,b), which coincides with the location of a strong CAPE gradient on the observations (Fig. 10c). The formation of a dry line can often be a precursor for severe thunderstorm formation with tornadogenesis potential (Grazulis, 2001). Note however, that the model missed the location of the core of Maximum CAPE (~3500 J kg^{-1}) by one degree, latitude and longitude. The observed maximum CAPE was located over the ocean, just off the Mississippi coast, while in the model output; the same core was placed at the southern Mississippi-Louisiana border. Nonetheless, by assimilating lightning flash rates, the analysis increased, thus increasing the magnitude of winds and absolute vorticity at 850 hPa. The analysis increment of wind, suggests that absolute vorticity was advected into the region of strong CAPE gradient (dry-line).

4.4 Statistics: analysis and forecast Root Mean Square (RMS) errors with respect to the lightning observations (LIGHT vs. NODA)

A qualitative comparison of atmospheric fields between the data assimilation (LIGHT) and the control (NODA) experiments with observations may lead to subjective conclusions on determining which experiment outperformed the other. Statistical evaluations on the other hand, can provide useful diagnostics when morphological differences are not obvious.

Analysis and forecast RMS errors with respect to the lightning observations were calculated from a domain containing the observed lightning flash rates at 10 km resolution during the 6-hour assimilation time window, as described in Sect. 2.2. From Fig. 11a, the LIGHT experiment achieves a better fit in the analysis compared to the NODA experiment, but not for cycle 6. A possible reason could be that the system was exiting the model domain at that time. Since the strongest convection and cold front moved away from the domain, there was no significant lightning activity over the region. Consequently, the number of lightning observations available for data assimilation significantly decreased and the impact of lightning data assimilation was reduced. The analysis result is not well retained in the forecast (Fig. 11b). This issue definitely requires further investigation. A possible reason may be that there are no other types of observations being assimilated, such as conventional and satellite observations that would additionally constrain the analysis and eventually create dynamical balance, further improving the analysis and consequently the forecast. Note that lightning is just an additional type of observation. All available observations have to be in agreement with each other at the same location. Therefore, in regions where lightning observations are not in agreement with other types of observations, the data assimilation algorithm will create the optimal observation impact based on uncertainty of all observations in the region. In areas where lightning observations are not available other measurements should help.

5 Summary and future work

In this study, the preliminary development and assessment of a methodology for the assimilation of lightning observations through hybrid variational-ensemble methods is presented. The aim of the study was to evaluate if lightning data assimilation can be useful in mesoscale, regional, and global applications at a coarse resolution in which convection cannot be explicitly resolved. The MLEF system interfaced with WRF-NMM was utilized to investigate the impacts of lightning data assimilation on a mesoscale NWP model. As a proof of concept, this methodology was tested for the 27-28 April 2011 severe weather event in the southeastern United States. Results indicate that lightning was capable of spreading new information into the WRF-NMM model. Analysis increments of 750 hPa specific humidity, temperature, and winds indicate that the assimilation of lightning flash rates could impact the initial conditions of a subset of model variables (*q, T, U* and *V*) leading to dynamical balance as shown by the output from the 1-OBS test. The information content of lightning data was quantified through the calculation of degrees of freedom for signal. Regions of high density of observed lightning flash rates were in agreement with information content theory indicating that the time-flow-dependent forecast error covariance was directly related to observations during the assimilation period.

Evaluation of some atmospheric fields from the LIGHT experiment indicated that the assimilation of lightning data influenced winds, absolute vorticity and CAPE. A core of increased background wind speed at 850 hPa coincides with the location of the region of high density in lightning observations for the same assimilation cycle, indicating that the assimilation of lightning data had an impact on the increase of wind speed. Analysis increments of the 850 hPa wind, absolute vorticity and background CAPE indicated that vorticity was advected into the region of strong CAPE gradient where a dry-line formed. All these changes suggest the development of an environment favorable for strengthening of deep convection.

Analyses and forecast RMS errors with respect to the lightning observations from the LIGHT and NODA experiments indicated that LIGHT achieved a better fit at the analysis step compared to the NODA experiment. However, the 6-hour forecast after assimilation did not show any clear improvements in terms of the RMS errors. This requires further investigation.

The methodology presented in this study represents an initial step towards developing a comprehensive multivariate, multi-scale, multi-sensor operational data assimilation system that prepares for the assimilation of lightning along with different types of operational observations and for multiple applications. As a first step, we intended to verify if the data assimilation techniques described here could be accomplished and that lightning data could add information content to a modeling system with a coarse resolution similar to the ones used in operations. Further studies are planned where this methodology will be tested for different applications (e.g. different case studies, different models, and choice of observation operators). Operational conventional and satellite observations will be assimilated alongside lightning flash rates to further constrain the fit in the analysis.

6 Acknowledgements

This work was supported by the NOAA GOES-R Risk Reduction Program Grant NA17RJ1228 (Amendment 175), and the Joint Center for Satellite Data Assimilation program Grant NA10NES4400012. This work was also partially supported by the National Science Foundation Collaboration in Mathematical Geosciences Grant ATM-0930265. We would like to acknowledge the high-performance computing support provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation. Data, valuable input and some plots were provided by the following individuals:

Daniel Bikos, Gregory DeMaria, Robert DeMaria, Gaël Descombes, Jack Dostalek, and Daniel Lindsey.

Disclaimer: The views, opinions, and findings contained in this article are those of the authors and should not be construed as an official National Oceanic and Atmospheric Administration (NOAA) or U.S. Government position, policy, or decision.

7 Appendix. Lightning flash rate observation operator correction: Weak Constrain

Assume a multiplicative correction to the observation operator (i.e. correction in magnitude, not in the direction of the vector)

$$h(x) \rightarrow \alpha h(x), \tag{A1}$$

where $\alpha > 0$ is the unknown multiplication parameter.

Consider a logarithmic function of vectors since all vectors (i.e. y and $h(x)$) are positive definite and define a cost function with the adjustable parameter α:

$$\begin{aligned} J(\alpha) &= \frac{1}{2}\left[\log(\alpha) - \log(\alpha_0)\right]^T \mathbf{W}^{-1}\left[\log(\alpha) - \log(\alpha_0)\right] \\ &+ \frac{1}{2}\left[\log(y) - \log(\alpha h(x^f))\right]^T \mathbf{R}_L^{-1}\left[\log(y) - \log(\alpha h(x^f))\right] \end{aligned} \tag{A2}$$

where $\mathbf{R}_L$ is the observation error covariance associated with a logarithmic transformation, α_0 is a guess value, and $\mathbf{W}$ is the uncertainty matrix of the guess value. The optimal parameter $\alpha_{opt} > 0$ that minimizes the cost function (A2) is searched for. Following a standard procedure of function minimization to solve:

$$\left(\frac{\partial J(\alpha)}{\partial \alpha}\right)_{\alpha_{opt}} = 0 \quad . \tag{A3}$$

Note that in order to differentiate with respect to *a* it may be more convenient to redefine the cost function (A2) in the following manner:

$$\begin{aligned} J(\alpha) &= \frac{1}{2}\left[\log(\alpha) - \log(\alpha_0)\right]^T W^{-1}\left[\log(\alpha) - \log(\alpha_0)\right] \\ &+ \frac{1}{2}\left[\log\left(\frac{y}{h(x)}\right) - \log(\alpha)\right]^T R_L^{-1}\left[\log\left(\frac{y}{h(x)}\right) - \log(\alpha)\right] \end{aligned} \quad . \tag{A4}$$

The Jacobian of (A4) is

$$\frac{\partial J(\alpha)}{\partial \alpha} = \frac{1}{\alpha}[1]^T W^{-1}\left[\log(\alpha) - \log(\alpha_0)\right] - \frac{1}{\alpha}[1]^T R_L^{-1}\left[\log\left(\frac{y}{h(x)}\right) - \log(\alpha)\right], \qquad \text{(A5)}$$

where $[\boldsymbol{1}]$ is a vector with all components equal to one. After employing (A3)

$$\frac{1}{\alpha}\left\{(\log(\alpha))[1]^T W^{-1}[1] + \log\alpha[1]^T R_L^{-1}[1] - [1]^T R_L^{-1}\left[\log\left(\frac{y}{h(x)}\right)\right] - \log(\alpha_0)[1]^T W^{-1}[1]\right\} = 0\,. \qquad \text{(A6)}$$

After multiplying (A6) by α (where $\alpha > 0$) (A6) can be rewritten as

$$(\log(\alpha))[1]^T\left[R_L^{-1} + W^{-1}\right][1] - [1]^T R_L^{-1}\left[\log\left(\frac{y}{h(x)}\right)\right] - \log(\alpha_0)[1]^T W^{-1}[1] = 0\,. \qquad \text{(A7)}$$

From (A7):

$$\log(\alpha) = \frac{\left([1]^T R_L^{-1}\left[\log\frac{y}{h(x)}\right] + \log(\alpha_0)[1]^T W^{-1}[1]\right)}{[1]^T\left[R_L^{-1} + W^{-1}\right][1]}\,. \qquad \text{(A8)}$$

Finally, the optimal multiplicative parameter is given by:

$$\alpha_{opt} = \exp\left[\frac{[1]^T R_L^{-1}\left[\log\left(\frac{y}{h(x)}\right)\right] + \log(\alpha_0)[1]^T W^{-1}[1]}{[1]^T\left[R_L^{-1} + W^{-1}\right][1]}\right] \qquad \text{(A9)}$$

After employing a common assumption that the uncertainty matrix **W** and the observation error matrix **R**$_L$ are diagonal, with $diag(\mathbf{W}) = w_0$ and $diag(\mathbf{R}_L) = r_0$, respectively,

$$[1]^T W^{-1}[1] = N_{obs} w_0^{-1} \qquad \text{(A10)}$$

$$[1]^T \left[R_L^{-1} + W^{-1}\right][1] = N_{obs}\left(r_0^{-1} + w_0^{-1}\right) \qquad \text{(A11)}$$

$$[1]^T R_L^{-1}\left[\log\left(\frac{y}{h(x)}\right)\right] = r_0^{-1}\sum_{i=1}^{N_{obs}}\left[\log\left(\frac{y}{h(x)}\right)\right]_i . \qquad \text{(A12)}$$

where N_{obs} is the number of observations. By substituting (A10), (A11), and (A12) in (A9) gives:

$$\alpha_{opt} = \exp\left\{\frac{\frac{1}{N_{obs}}\sum_{i=1}^{N_{obs}}\log\left(\frac{y}{h(x)}\right)_i + \left(\frac{w_0}{r_0}\right)^{-1}\log(\alpha_0)}{1+\left(\frac{w_0}{r_0}\right)^{-1}}\right\}. \qquad \text{(A13)}$$

Without additional knowledge, a typical guess value is α_0 = 1, which further simplifies the solution (A13) to

$$\alpha_{opt} = \exp\left[\frac{\frac{1}{N_{obs}}\sum_{i=1}^{N_{obs}}\log\left(\frac{y}{h(x)}\right)_i}{1+\frac{r_0}{w_0}}\right]. \qquad \text{(A14)}$$

The above expression can be easily calculated in the observation operator and provide an adjustable correction factor.

References

Adamo C,. Goodman S., Mugnai A., Weinman J. A.: Lightning measurements from satellites and significance for storms in the Mediterranean. Betz H. D., Schumann U., Laroche P. (Eds.), Lightning Principles, Instruments and Applications: Review of Modern Lightning Research, Springer: Dordrecht, pp 309-329, 2009.

Alexander G. D., Weinman J. A., Karyampudi V., Olson W. S., Lee A. C. L.: The effect of assimilating rain rates derived from satellites and lightning on forecasts of the 1993 Superstorm. Mon. Weather Rev. 127, 1433–1457, 1999.

Barthe C., Deierling W., Barth M. C.: Estimation of total lightning from various storm parameters: a cloud resolving model study. J. Geophys. Res. 115, D24202, DOI:10.1029/2010JD014405, 2010.

Black R. A., Hallett J.: Electrification of the hurricane. J. Atmos. Sci. 56: 2004-2028, 1999.

Ek M. B., Mitchell K. E., Lin Y., Rogers E., Grunmann P., Koren V., Gayno G., Tarpley J. D.: Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. J. Geophys. Res. 108, (D22), 8851, DOI:10.1029/2002JD003296, 2003.

Falkovich A., Ginis I., Lord S.: Ocean data assimilation and initialization procedure for the Coupled GFDL/URI Hurricane Prediction System. J. Atmos. Ocean. Tech. 22, 1918-1932, 2005.

Ferrier B. S.: An efficient mixed-phase cloud and precipitation scheme for use in operational NWP models. Eos, Trans. Amer. Geophys. Union, 86(18), Jt. Assem. Suppl., A42A-02, 2005.

Fierro A. O., Mansell E. R., Ziegler C. L., MacGorman D. R.: Application of a lightning data assimilation technique in the WRF-ARW model at cloud-resolving scales for the tornado outbreak of 24 May 2011. Mon. Weather Rev. 140, 2609-2627, 2012.

Finke U.: Optical detection of lightning from space. Betz H. D., Schumann U., Laroche P. (Eds.), Lightning Principles, Instruments and Applications: Review of Modern Lightning Research, Springer, Dordrecht, pp 271-286, 2009.

Fletcher, S. J. and Zupanski M.: A Data Assimilation Method for Lognormally Distributed Observational Errors. Q.J. Roy. Meteorol. Soc., **132**, 2505-2519, 2006.

Grazulis T. P.: The Tornado: Nature's Ultimate Windstorm. University of Oklahoma Press: Norman, pp 324, 2001.

Hayes J. L.: U.S. Department of Commerce, National Oceanic and Atmospheric Administration, National Weather Service. Service assessment: The historic tornadoes of April 2011. Silver Spring, Maryland. Accessed 3 Nov 2012, “available at: http://www.nws.noaa.gov/os/assessments/pdfs/historic_tornadoes.pdf, 2011.”

Janjić Z. I.: The Step-Mountain Eta Coordinate Model: Further developments of the Convection, Viscous Sublayer, and Turbulence Closure schemes. Mon. Weather Rev. 122, 927–945, 1994.

Janjić Z. I.: Comments on ”Development and evaluation of a convection scheme for use in climate models, J. Atmos. Sci. 57, 3686, 2000.

Janjić Z., Black T., Pyle M., Rogers E., Chuang H. Y., DiMego G. High resolution applications of the WRF NMM. In Proceedings of 21st Conf. on Weather Analysis and Forecasting/17th Conf. on Numerical Weather Prediction, Washington, DC, 31 July – 5 August 2005, American Meteorological Society, Boston, 2005.

Janjić Z., Gall R., Pyle M. E.: Scientific documentation for the NMM solver, Technical Note TN-477+STR, 53, 125pp. NCAR, Boulder, Colorado, USA, 2010.

Lay E. H., Holzworth R. H., Rodger C. J., Thomas J. N., Pinto O., Dowden R. L. WWLLN global lightning detection system: regional validation study in Brazil. Geophys. Res. Lett. 31, L03102, DOI:10.1029/2003GL018882, 2004.

Lien, G.-Y., Kalnay E., and Miyoshi T.: Effective assimilation of global precipitation: Simulation experiments. Tellus *A*, 65, 19915. doi:10.3402/tellusa.v65i0.19915, 2013.

Mansell E. R., Ziegler C. L., MacGorman D. R.: A lightning data assimilation technique for mesoscale forecast models. Mon. Weather Rev. 135, 1732–1748, 2007.

Papadopoulos A., Chronis T. G., Anagnostou E. N.: Improving convective precipitation forecasting through assimilation of regional lightning measurements in a mesoscale model. Mon. Weather Rev. 133, 1961–1977, 2005.

Pessi A. T. and Businger S.: The impact of lightning data assimilation on a winter storm simulation over the North Pacific Ocean. Mon. Weather Rev. 137, 3177-3195, 2009.

Price C. G. and Rind D.: A simple lightning parameterization for calculating the global lightning distribution. J. Geophys. Res. 97: 9919-9933, 1992.

Price C. G.: Thunderstorms lightning and climate change. Betz H. D., Schumann U., Laroche P. (Eds.), Lightning Principles, Instruments and Applications: Review of Modern Lightning Research, Springer, Dordrecht, pp 521-535, 2009.

Price C. G.: Lightning applications in weather and climate research. Surv. Geophys. 34:6, 755-767, 2013.

Rodgers C. D.: Inverse Methods for Atmospheric Sounding: Theory and Practice. World Scientific, Singapore, pp 256, 2000.

Shannon C. E., Weaver W.: The Mathematical Theory of Communication. University of Illinois Press, Champaign, pp. 144, 1949.

Zupanski M.: Maximum likelihood ensemble filter: theoretical aspects. Mon. Weather Rev. 133, 1710-1726, 2005.

Zupanski D., Hou A. Y., Zhang S. Q., Zupanski M., Kummerow C. D., Cheung S. H.: Applications of information theory in ensemble data assimilation. Q. J. Roy. Meteor. Soc. 133, 1533-1545, 2007.

Zupanski M., Navon I. M., Zupanski D.: The maximum likelihood ensemble filter as a non-differentiable minimization algorithm. Q. J. Roy. Meteor. Soc. 134, 1039-1050, 2008.

Figures

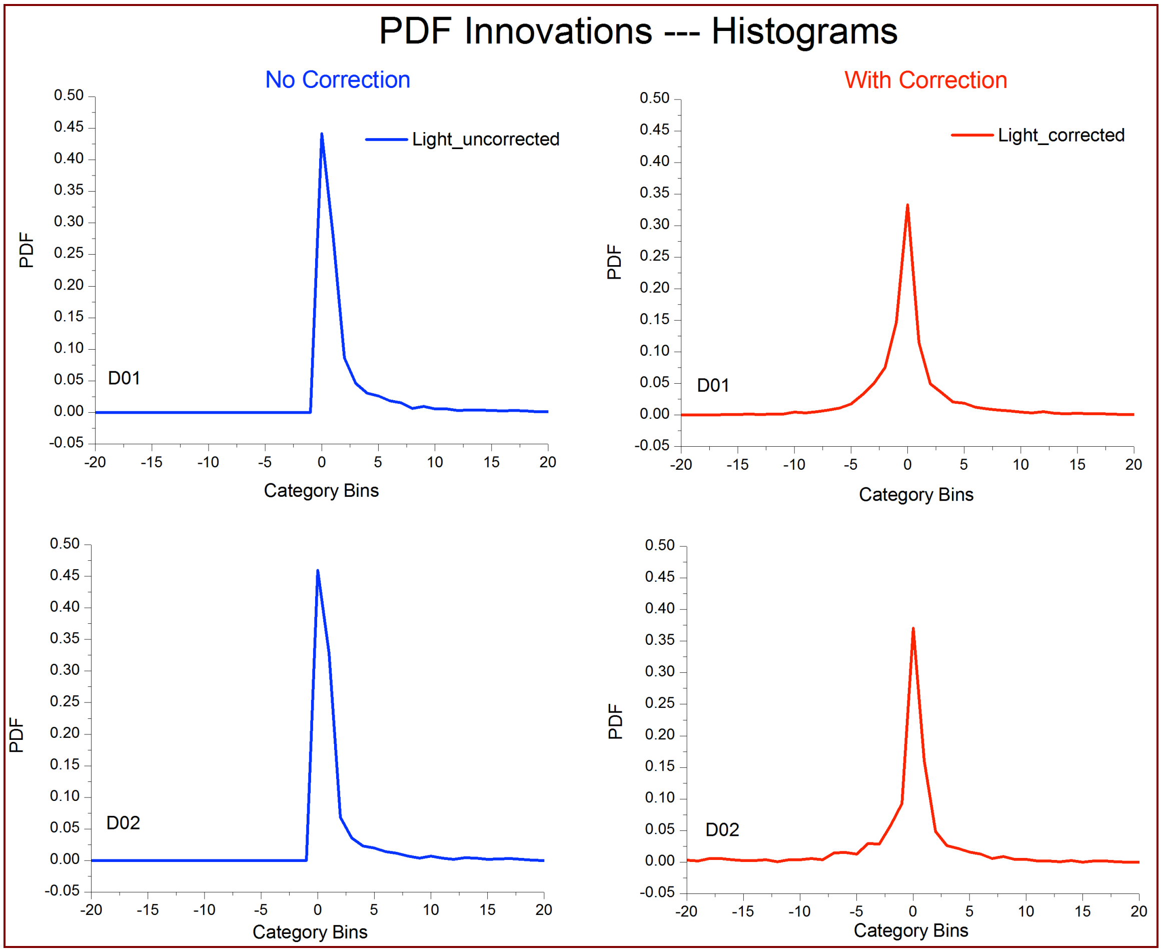

Fig. 1. Statistics of normalized innovation vectors $R^{-1/2}\left[y - h(x^f)\right]$, or PDF innovations for cycles 1-5 for both domains (D01 and D02) before (left-blue) and after (right-red) correction. The skewed histograms on the left implicitly indicate that the values of observed lightning flash rate are considerably larger than the guess, a situation that required a correction.

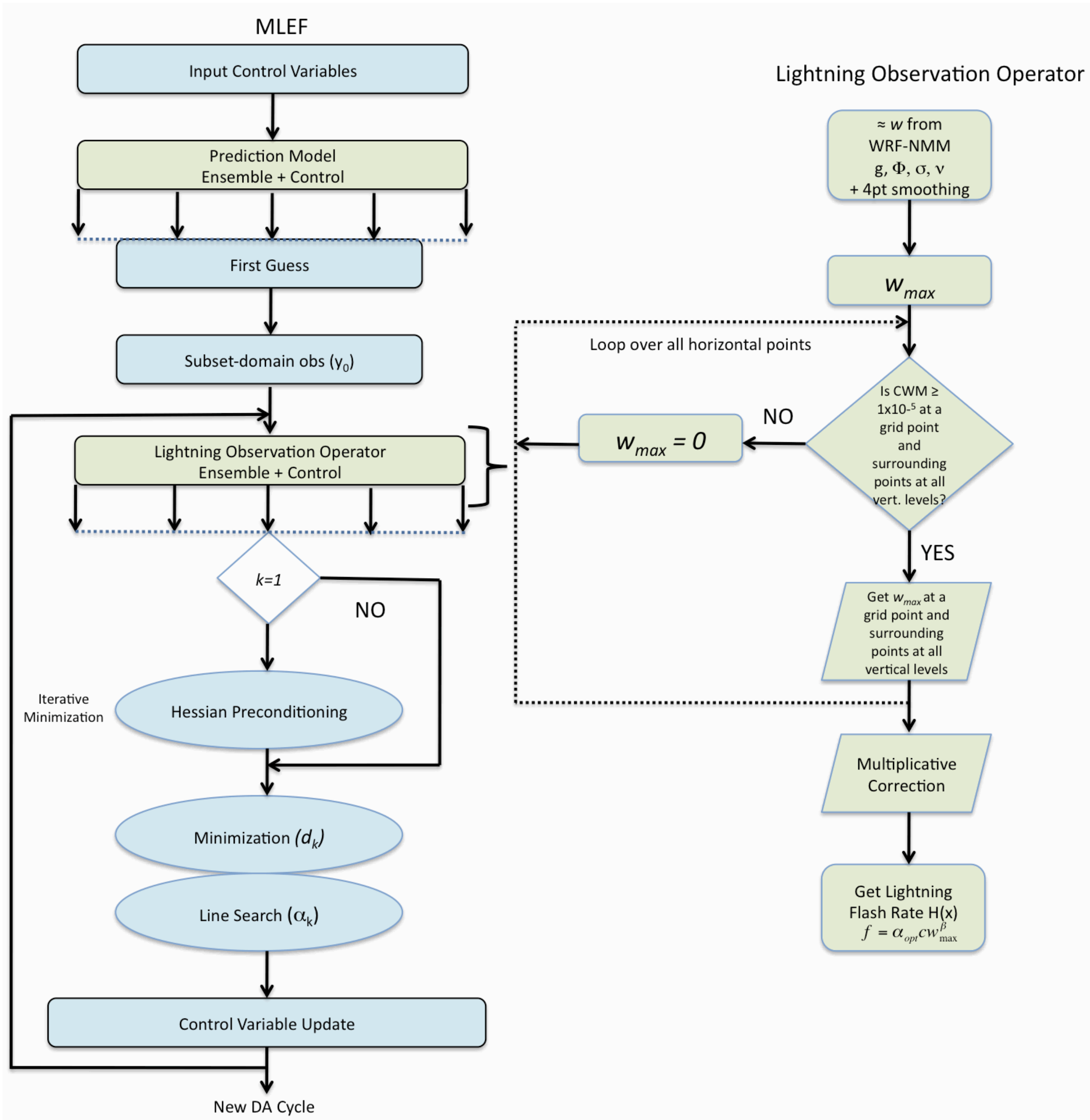

Fig. 2. Flow chart of the data assimilation system, the left section is the MLEF system with all its components. The lightning observation operator algorithm is shown on the right-hand side of the flow chart.

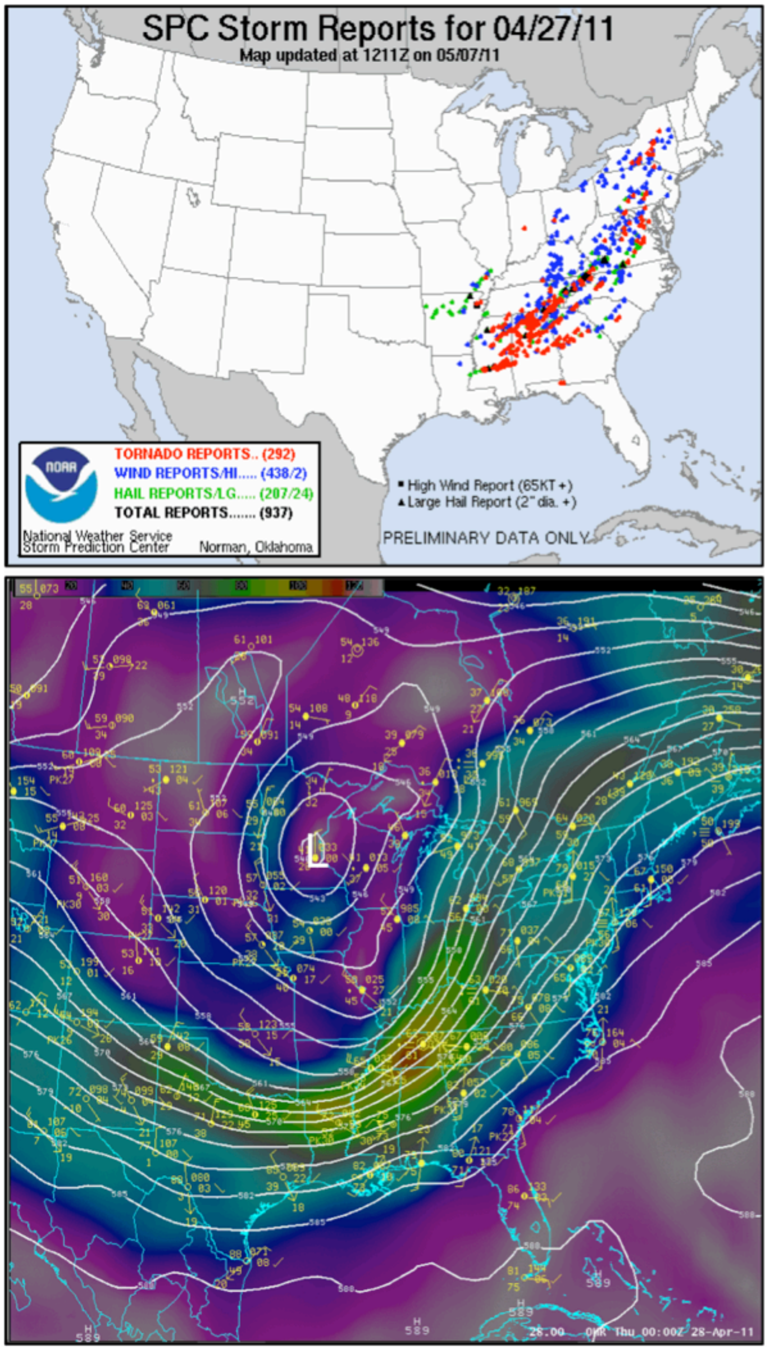

Fig. 3. Valid at 0000 UTC 28 April 2011. Storm Prediction Center daily storm reports showing a total of 292 reported tornados (top). Forecast Systems Laboratory, 500 hPa geopotential heights and color contoured wind, and surface observations (bottom), showing an upper level low over Minnesota, a deep trough with an associated jet streak over the northeastern corner of Alabama, indicative of a region of positive vorticity adevection (bottom, Courtesy of Daniel Bikos).

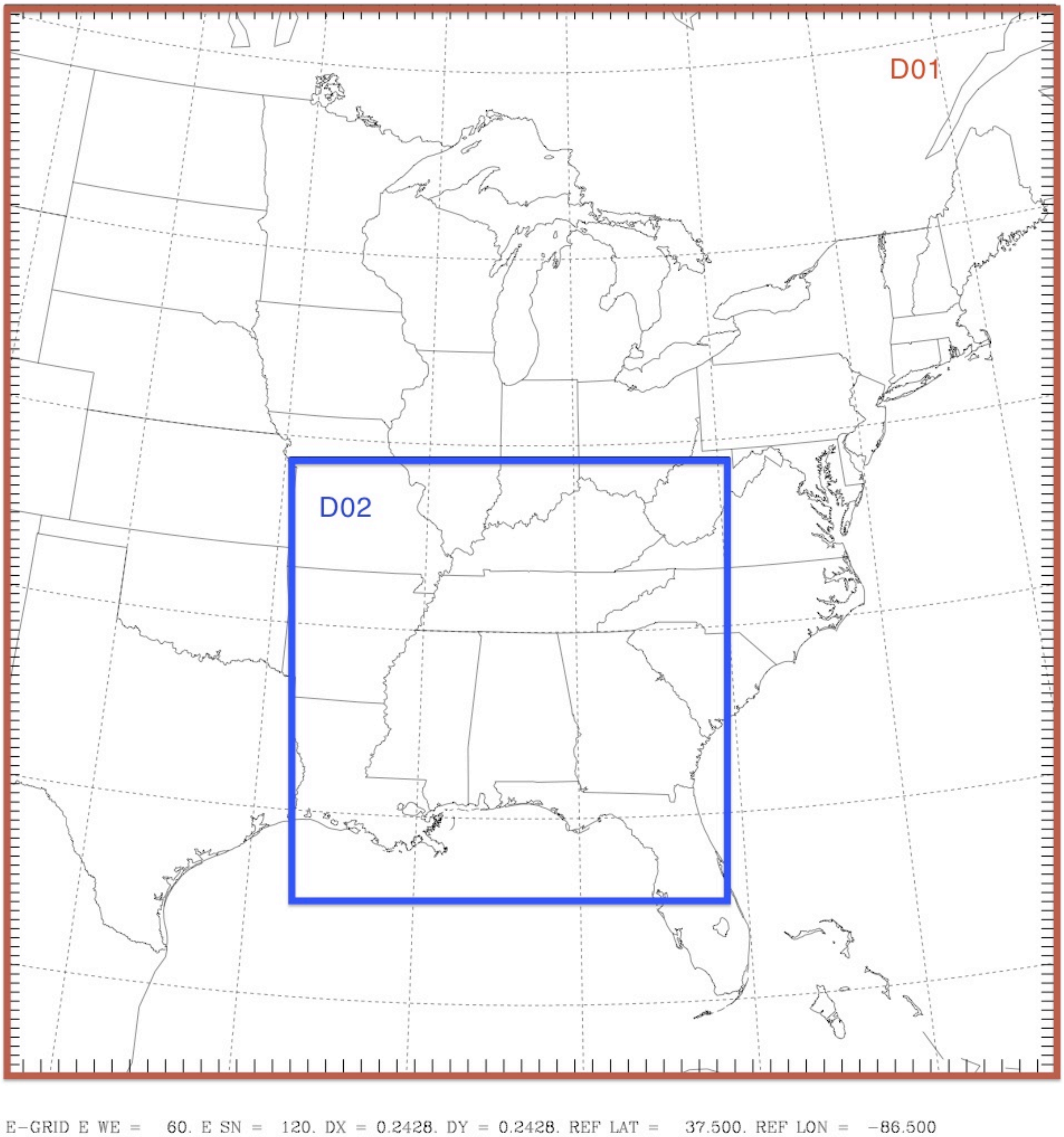

Fig. 4. Domain configuration. D01 is the mother domain with a size of 1350 by 2952 km^2 (50 x 96 grid points) at 27 km resolution. D02, the inner nest has a size of 540 by 1170 km^2 (60 x 130 grid points) at 9 km resolution.

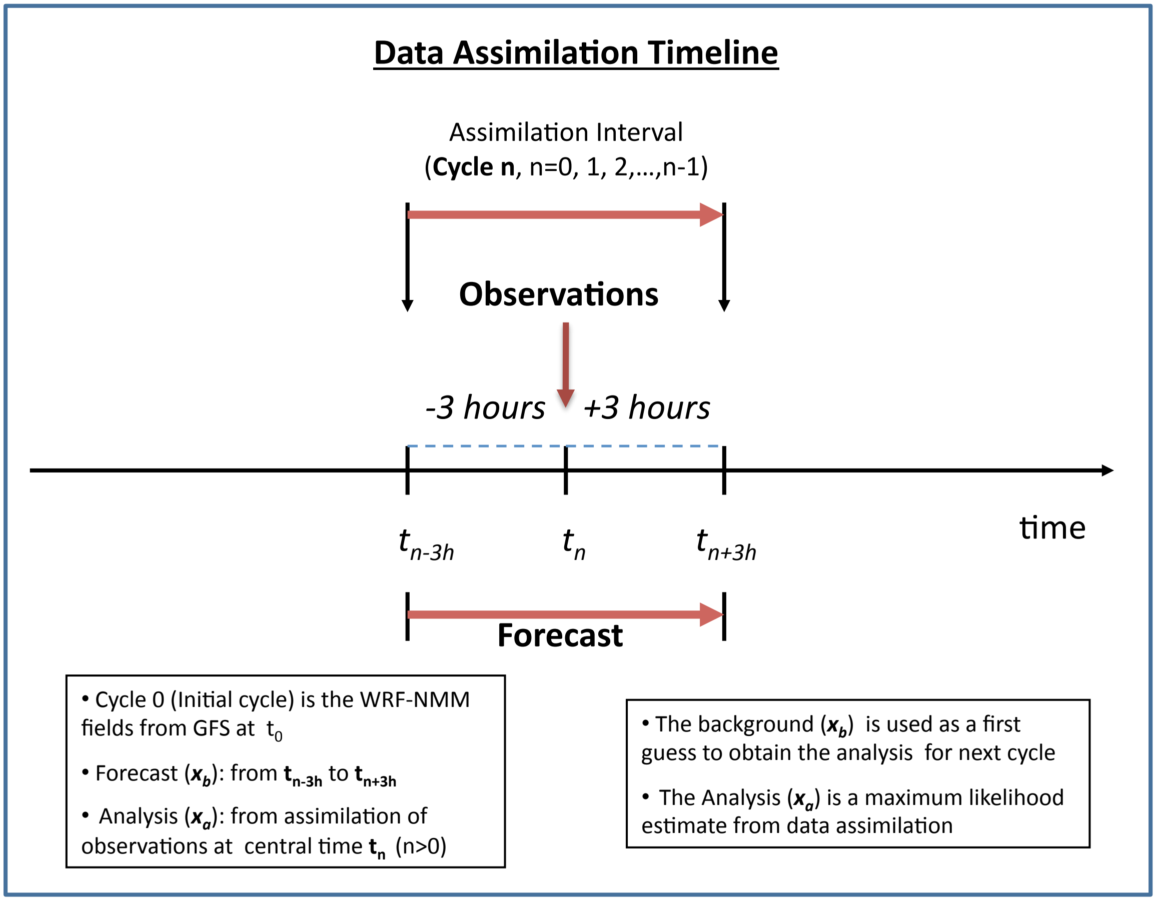

Fig. 5. Data assimilation timeline, the data assimilation frequency for the lightning observation is 6 hours (±3 hours) from a central time t_n>0. The initial cycle (Cycle 0) is just the model (WRF-NMM) output fields from the GFS files, at t_n, the forecast, or background state ($\boldsymbol{x_b}$) is obtained from t_{n-3h} to t_{n+3h}. The forecast is used as a guess to obtain the analysis solution for the next cycle.

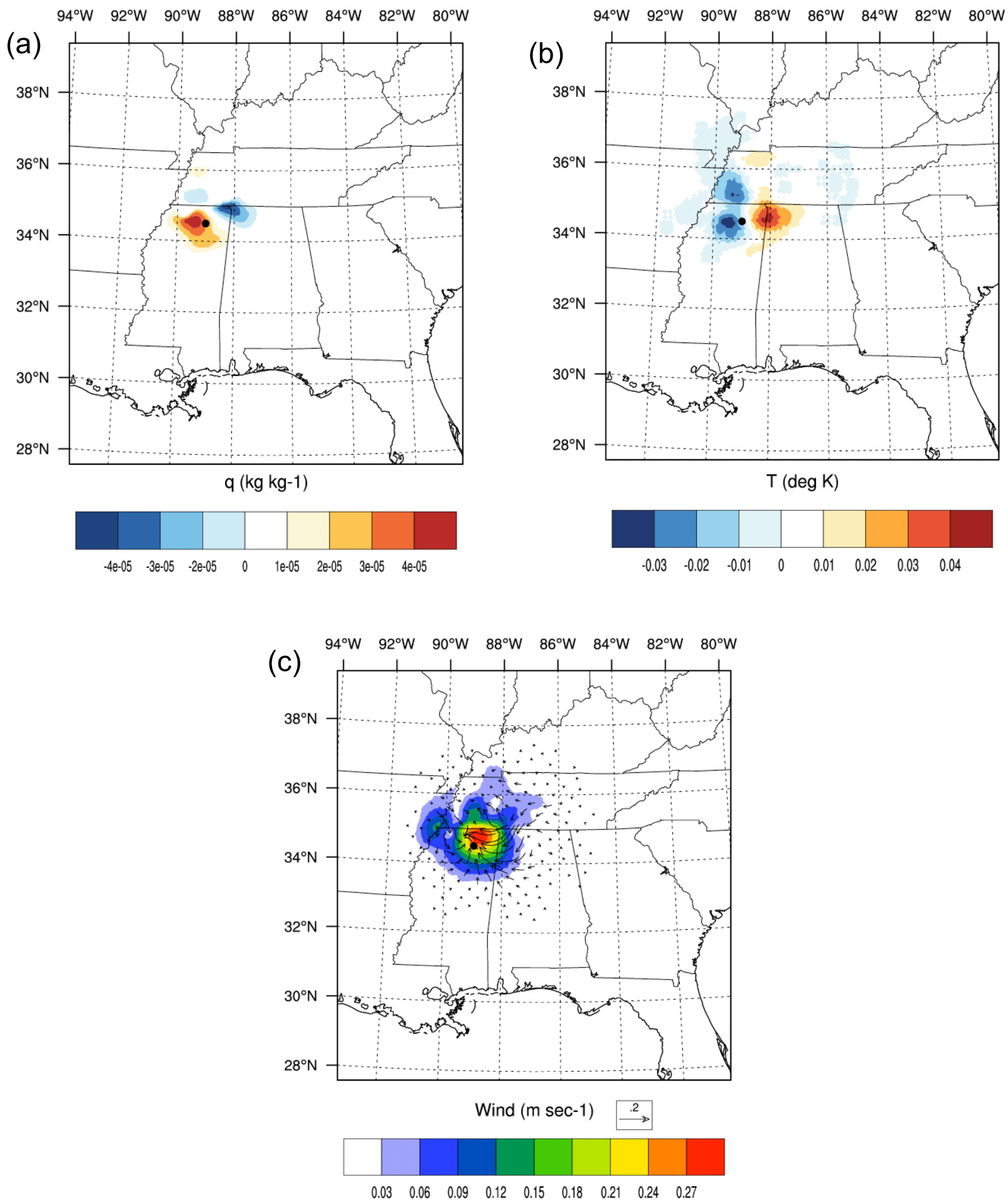

Fig. 6. Analysis increments of **(a)** specific humidity, **(b)** temperature and **(c)** wind at 700 hPa. The black dot shows the location of the single observation (35.01°N, 87.60°W). Dipoles of positive and negative analysis increments can be observed at either end of the single observation in the specific humidity and temperature plots, but with opposite signs. 700 hPa winds show a positive analysis increment with maximum values coinciding with the region of positive temperature increment and anti-cyclonic circulation can be observed around the location of the single observation.

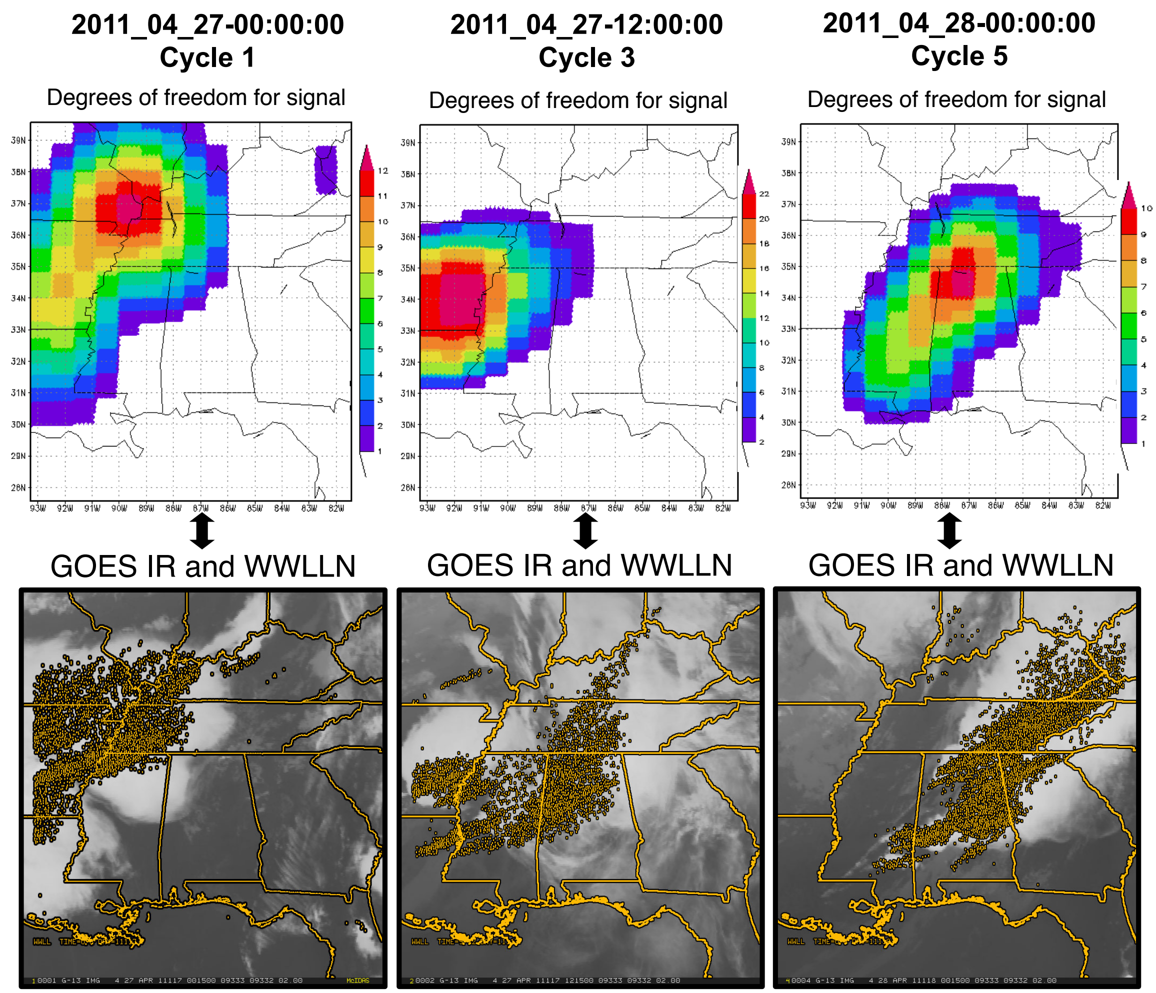

Fig. 7. Degrees of freedom for signal (top-three plots) of assimilated lightning data and observed GOES IR and WWLLN lightning flash rates (bottom-three plots, courtesy of Gregory DeMaria and Jack Dostalek) for cycles 1, 3, and 5. The areas of highest density of lightning observations are in general agreement with information content, implying that the flow-dependent ensemble forecast error covariance geographically coincides with throughout most of the assimilation period. Note, however, that the agreement for cycle 3 is not very good, implicitly confirming that ensemble forecast uncertainty is not always sufficient to represent the true forecast uncertainty.

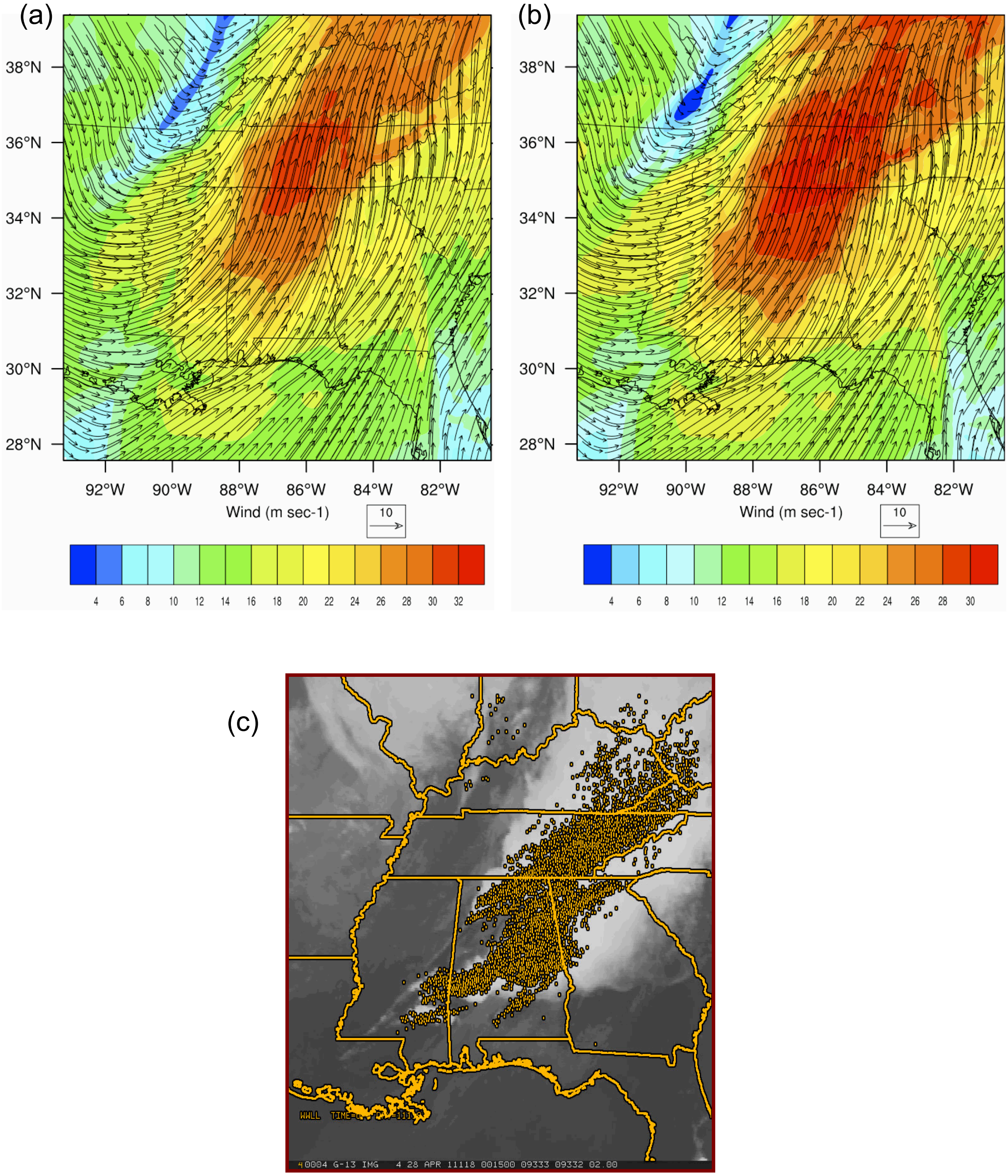

Fig. 8. (a) Background (forecast) winds at 850 hPa at 0000 UTC 28April 2011 (cycle 5) from the experiment without lightning (NODA), **(b)** background (forecast) winds at 850 hPa at 0000 UTC 28April 2011 (cycle 5) from the lightning data assimilation experiment (LIGHT) and **(c)** GOES IR and observed 6-hour WWLLN lightning flash rates at the same time (Courtesy of Gregory DeMaria and Jack Dostalek). The core of strong wind speed matches the region of high lightning flash rate density in the observations, but note that the core of maximum wind speed has a larger spatial coverage in the LIGHT experiment (b) and based on computed differences, stronger winds in the order of 4 m sec^{-1} were found in the LIGHT experiment.

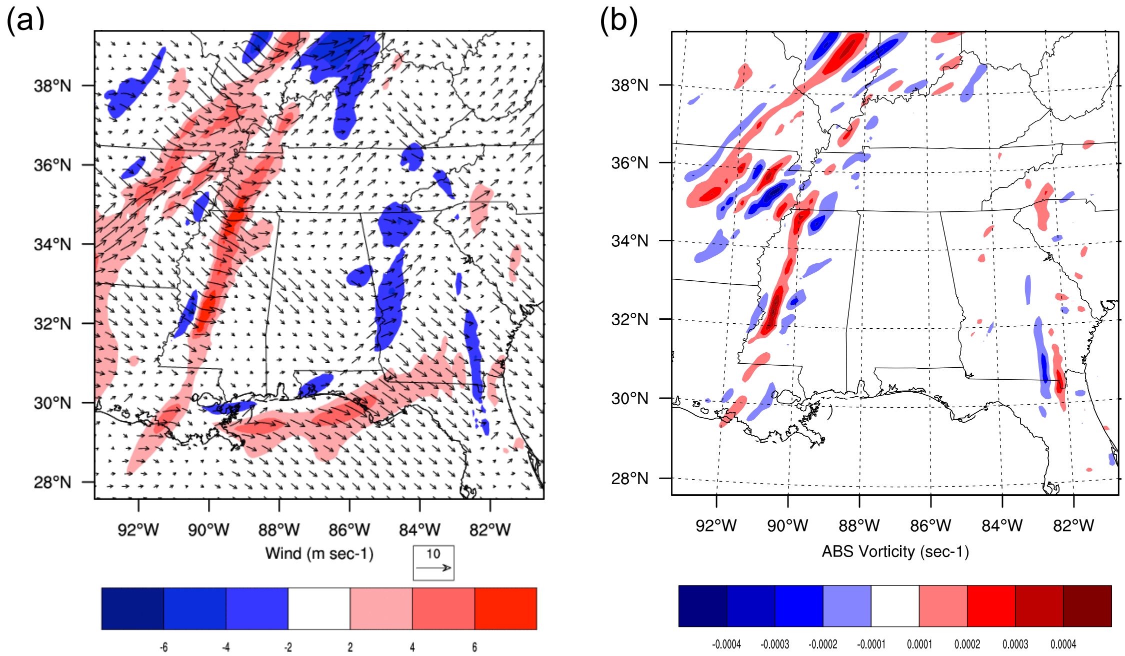

Fig. 9. Analysis increments at 850 hPa of **(a)** wind and **(b)** absolute vorticity at 0000 UTC 28 April 2011. Regions of positive increments are found in the upper left-hand side in both plots indicated by the ellipses. Winds are being advected into the region of strong CAPE seen in Fig. 10a.

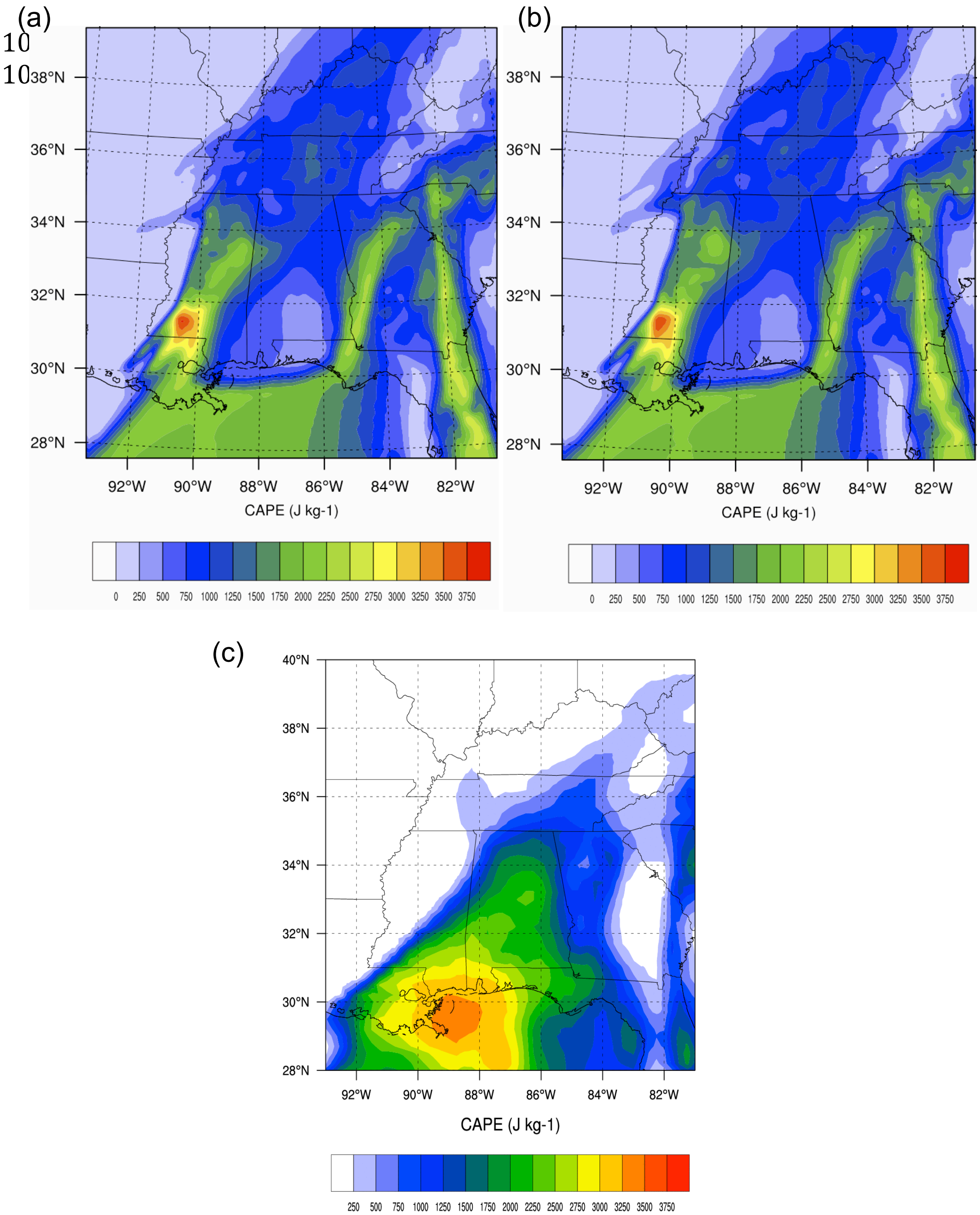

Fig. 10. Background CAPE for **(a)** NODA and **(b)** LIGHT experiments, and **(c)** observed CAPE from the Storm Prediction Center's Surface Mesoanalysis at 0000 UTC 28 April 2011 (cycle5). A region of high CAPE gradient is observed in the upper-left hand side of the domain, indicating the presence of a well-defined dry line, in agreement with observations, but there are no significant differences between both experiments. One reason is that there are no lightning observations in the region where the strongest CAPE was observed. Lightning data was not able to impact CAPE.

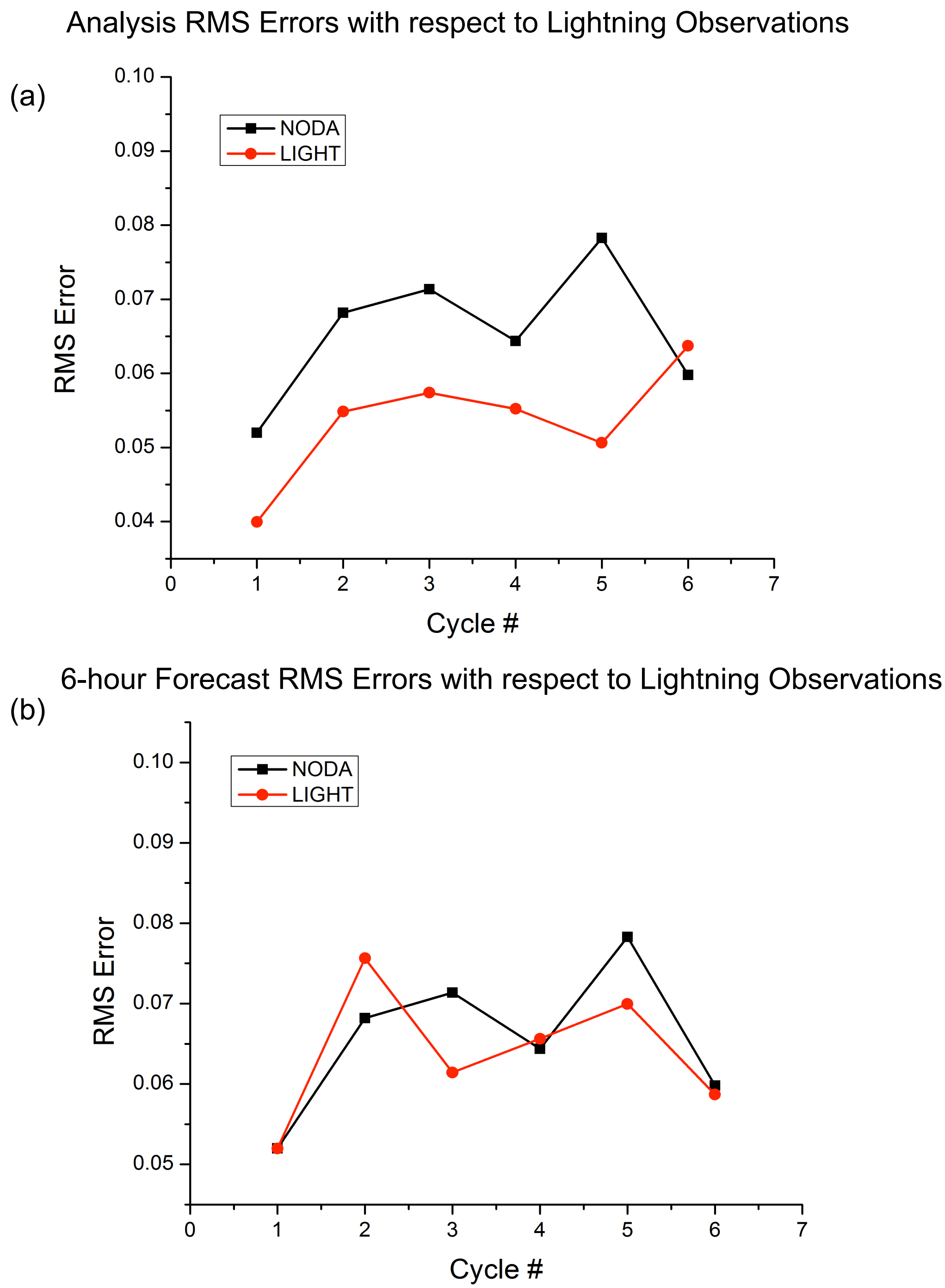

Fig. 11. Root mean square (RMS) errors with respect to lightning flash rate observations during six assimilation cycles at 6 h intervals: **(a)** Analysis RMS error. The RMS error reduction was achieved during the first 5 cycles of the assimilation period, while there is deterioration in the last cycle, possibly due to the fact that the system was exiting the model domain. **(b)** 6 h forecast RMS error. There is no clear improvement in the forecast, suggesting that additional development of the assimilation system might be required, such as an

improvement of the observation operator, adding new observations, and possibly improving the forecast uncertainty estimation.

Figure Captions

Fig. 1. Statistics of normalized innovation vectors $R^{-1/2}\left[y - h(x^f)\right]$, or PDF innovations for cycles 1-5 for both domains (D01 and D02) before (left-blue) and after (right-red) correction. The skewed histograms on the left implicitly indicate that the values of observed lightning flash rate are considerably larger than the guess, a situation that required a correction.

Fig. 2. Flow chart of the data assimilation system, the left section is the MLEF system with all its components. The lightning observation operator algorithm is shown on the right-hand side of the flow chart.

Fig. 3. Valid at 0000 UTC 28 April 2011*.* Storm Prediction Center daily storm reports showing a total of 292 reported tornados (top). Forecast Systems Laboratory, 500 hPa geopotential heights and color contoured wind, and surface observations (bottom), showing an upper level low over Minnesota, a deep trough with an associated jet streak over the northeastern corner of Alabama, indicative of a region of positive vorticity adevection (bottom, Courtesy of Daniel Bikos).

Fig. 4. Domain configuration. D01 is the mother domain with a size of 1350 by 2952 km^2 (50X96 grid points) at 27 km resolution. D02, the inner nest has a size of 540 by 1170 km^2 (60X130 grid points) at 9 km resolution.

Fig. 5. Data assimilation timeline, the data assimilation frequency for the lightning observation is 6 hours (±3 hours) from a central time t_n>0. The initial cycle (Cycle 0) is simply the model (WRF-NMM) output fields from the GFS files, at t_n, the forecast, or background state ($\boldsymbol{x_b}$) is obtained from t_{n-3h} to t_{n+3h}. The forecast is used as a guess to obtain the analysis solution for the next cycle.

Fig. 6. Analysis increments of **(a)** specific humidity, **(b)** temperature and **(c)** wind at 700 hPa. The black dot shows the location of the single observation (35.01°N, 87.60°W). Dipoles of positive and negative analysis increments can be observed at either end of the single observation in the specific humidity and temperature plots, but with opposite signs. 700 hPa winds show a positive analysis increment with maximum values coinciding with the region of positive temperature increment and anti-cyclonic circulation can be observed around the location of the single observation.

Fig. 7. Degrees of freedom for signal (top-three plots) of assimilated lightning data and observed GOES IR and WWLLN lightning flash rates (bottom-three

plots, courtesy of Gregory DeMaria and Jack Dostalek) for cycles 1, 3, and 5. The areas of highest density of lightning observations are in general agreement with information content, implying that the flow-dependent ensemble forecast error covariance geographically coincides with throughout most of the assimilation period. Note, however, that the agreement for cycle 3 is not very good, implicitly confirming that ensemble forecast uncertainty is not always sufficient to represent the true forecast uncertainty.

Fig. 8. **(a)** Background (forecast) winds at 850 hPa at 0000 UTC 28April 2011 (cycle 5) from the experiment without lightning (NODA), **(b)** background (forecast) winds at 850 hPa at 0000 UTC 28April 2011 (cycle 5) from the lightning data assimilation experiment (LIGHT) and **(c)** GOES IR and observed 6-hour WWLLN lightning flash rates at the same time (Courtesy of Gregory DeMaria and Jack Dostalek). The core of strong wind speed matches the region of high lightning flash rate density in the observations, but note that the core of maximum wind speed has a larger spatial coverage in the LIGHT experiment (b) and based on computed differences, stronger winds in the order of 4 m sec^{-1} were found in the LIGHT experiment.

Fig. 9. Analysis increments at 850 hPa of **(a)** winds and **(b)** absolute vorticity at 0000 UTC 28 April 2011. Regions of positive increments are found in the upper left-hand side in both plots. Winds are being advected to the region of strong CAPE seen in Fig. 10a.

Fig. 11. Root mean square (RMS) errors with respect to lightning flash rate observations during six assimilation cycles at 6 h intervals: **(a)** Analysis RMS error. The RMS error reduction was achieved during the first 5 cycles of the assimilation period, while there is deterioration in the last cycle, possibly due to the fact that the system was exiting the model domain. **(b)** 6 h forecast RMS error. There is no clear improvement in the forecast, suggesting that additional development of the assimilation system might be required, such as an improvement of the observation operator, adding new observations, and possibly improving the forecast uncertainty estimation.