<u>Review of nhess-2018-319:</u>

<u>Title</u>: "The impact of lightning and radar data assimilation on the performance of very short term rainfall forecast for two case studies in Italy." by Federico et al.

<u>Summary</u>: The authors utilize a cloud-scale functional relationship between lightning and water vapor mass mixing ratio published in the literature and applied it to a homegrown 3DVAR framework at the convection-allowing scale to evaluate the analysis and short term forecast of two selected high impact weather events over Italy.

<u>Recommendation:</u> reject and, eventually, re-submit.

<u>Main Comments:</u>

While the manuscript could eventually offer some merit for this journal, I found the analysis generally very rudimentary with the authors going at length in describing in excruciable level of details individual figures/panels in a repetitive and redundant manner without distilling the content into concise arguments/hypotheses. Given its repetitive nature, the entire results section could, in fact, easily be condensed into a 2-3 pages. Most importantly, the manuscript (hereafter, m/s) lacks rigor and rationales for the set ups and methods put forth for each, respective DA approaches. Salient Major issues are itemized below.

(1) As far as the scientific content is concerned, the core ideas and notions of this lightning data assimilation (LDA) method are conceptually similar to those from many existing studies, which fundamentally aim at promoting convective development through the introduction of latent heating within a prescribed neighborhood region/column centered at observed lightning locations. Past works from Benjamin et al. (2004), Alexander et al. (1999), Chang et al. (2001), Papadopoulos et al. (2005), Pessi and Businger (2009), have used empirical relationships between lightning-rainfall rates-latent heating or lightning-reflectivity rates-latent heating [e.g., in the HRRR]. Following a similar idea, recent works such as Machand and Fuelberg (2014), Lynn et al. (2015), Lynn (2017), Fierro et al. (2012; 2014, 2015), Wang et al. (2017, 2018) proposed LDA means that essentially boost the local thermal buoyancy where lightning is observed. A very limited portion of these techniques, however, offer alternative approaches to address spurious convection (i.e., removal) – which is a far more challenging problem to tackle. For completeness and given the relatively limited advances in LDA relative to radar DA, the authors should do a better job in discussing and including all the aforementioned references in their text. I was in fact astonished to notice that the integrity of the Results section in section 4 is completely devoid of references to previous works.
In particular, since they opted to borrow an LDA method from one of these investigators, comparisons with their study should be performed more systematically throughout the m/s. For instance, the works of Federico et al. 2017b is invoked when referring to multi-day forecast statistics using the Fierro et al. method without mentioning that, such a study, was already conducted by the same author over a larger domain and using nearly three times more forecast days/cases (Fierro et al. 2015 study). Given this omission, their study (Federico et al. 2017b) inadequately state that such multi-day statistics for this LDA have never been conducted. In a similar manner, it is of relevance to underline whenever appropriate that, in this work: (i) radial velocity is not included (specify why), (ii) only cloud-

to-ground lightning data are considered and (iii) spurious convection is not addressed. In the light of (i) and (ii), one on the recent studies they cite (Fierro et al. 2016) not only assimilated level II radar data (radial velocity + reflectivity factor) but used total lightning data. This needs to be clearly stated, for completeness (Cf comment 3 below for rationales).

(2). In term of DA methodology, I found one major drawback, which is never discussed, nor evaluated. Given that both the LDA and their "RAD" experiment make adjustments to the relative humidity (RH) field, it is expected that both techniques will overlap in their adjustments over all the (many) grid points characterized by observed lightning flash rates exceeding zero. This is because changing RH is equivalent to adjusting Qv as RH ~ Qv/Qv_saturation. A more self-consistent DA approach would adjust the pseudo-observations for the Qv or RH field in a manner that eliminates any possibility of overlap during the minimization. Toward that end, the authors should include soundings and/or horizontal cross sections of RH/Qv that shows, quantitively, how the RH field is adjusted by each respective DA approach (radar vs lightning).
Second, given that lightning is a cloud-scale observation, I cannot find any justifications for not conducting the 3DVAR analysis on the innermost, higher resolution domain. Instead, the method minimizes the cost function on the intermediate domain and, later, projects the innovations on the coarser-scale domain. This needs to be addressed.

Third, the radius of influence/decorrelation length scale chosen for radar reflectivity factor (50 km) is far too large for convective scale applications and would incur unrealistically large amount of Qv mass added into the domain – which will undoubtedly yield to spin-up issues and the generation of convective-scale gravity waves that will degrade longer term (>= 3h) solutions (please provide plot of perturbation pressure in your response). In that regard, the authors should indicate and contrast the total amount of Qv mass added by RAD and LIGHT.

(3). In the context of forecast improvements, the Qv-based method they borrowed/adapted was scaled for total lightning data (> 50% detection efficiency of intra-cloud [IC] flashes). I was surprised to find that absolutely no information on the detection efficiency and geolocation accuracy of the lightning network used (LINET) is provided in the text [no figures either]. Given the large area covered by this study, it is thus very likely that the geolocation accuracy of this network remains very poor for low amplitude flashes and for all flashes over oceanic regions. Given the low sferics amplitudes of IC flashes, the VLF portion of the sensor will miss nearly all these flashes, while the VHF portion only is able to detect some of the IC flashes within a few tens of kilometers away from the station [e.g., Rison, MacGorman works]. Thus, it is relevant to state and underscore that LINET only detects a very small portions of the total IC flashes in the study domain (likely < 5%). Motivation for scaling the F12 method for IC flashes (in lieu of cloud-to-ground [CG] flashes), lies in the well-documented finding that, in contrast to CGs, ICs are well correlated with thunderstorm kinematic and microphysical evolution (updraft strength, updraft volume, graupel mass etc, see Wiens et al. 2005, Schultz et al. 2011 among many others). CGs, on the other hand, were found to be correlated with the descent of reflectivity cores and the onset of the demise of the storm's updraft core [MacGorman and Nielsen 1991, MacGorman et al. 1989, Rutledge and Lang's seminal works etc]. Not surprisingly, ICs were found to lag CG by an average of 15 min [see one of the recent MacGorman study]. Moreover, Boccippio et al. 2001 and Medici et al. 2017 found that in deep continental convection, IC flashes always outnumber CGs by a ratio sometime exceeding 10:1. Based on these facts, it becomes clear why the Fierro method emphasized the use of IC flashes [or total lightning] for their application. Further motivation arises from the recent successful launches of the GLM instrument aboard GEOS-16/17, which will provide continuous day/night coverage of total lightning at ~90% detection efficiency (DE) over a large domain covering the Americas (Gurka et al. 2006; Goodman et al. 2012, 2013, Rudlosky et al. 2018). Note that GLM will provide flash extent information of lightning, while the metric derived from the (limited) point flash data in this study can only provide a very rough surrogate for CG flash location density at best. Similar space-borne technology to detect lightning have been developed by China (Feng-Yun-4, yang et al. 2016) with these data being assimilated in recent works by Wang et al. (2017, 2018) – which were never referenced either. Apart from their propensity to detect total lightning at a high DE, the chief advantage of this technology lies in its ability to retrieve lightning over remote oceanic regions.

(4) The following key information pertaining to the respective DA methods are missing/never discussed:

(a) What are the background/observation errors for reflectivity/lightning?

(b) What statistics are used for model error ?

(c) How is the adjoint for the lightning data assimilation operator derived ?

(d) What assumptions are made for grid points with zero lightning or zero reflectivity observations ? Does the radar DA or LDA treat those as missing observations or equate those to the background values to reduce spread ?

(e) What Gaussian decorrelation length scales are assumed for each observation ? Please specify/justify/explain. How would the selection of a given length scale value, influence the results ?

(f) Is spurious convection addresses by either DA method ? Please elaborate.

(g) Does the variational minimization set use a multi-scale approach ? If yes, what influence radii are chosen and how many cycles ?

(5) Why did the authors not include the fractions skill score FSS as the main evaluation metric for their forecast? Several works have posited that, in contrast to ETS, FSS does not penalize displacement errors as much and, arguably, FSS offers a more accurate measure of skill on convection-allowing grids (Mittermaier et al. 2011).

Additionally, more recent studies evaluating forecast performance have been making usage of the so-called performance diagrams, which conveniently merge several key contingency table elements into one single diagram (Roebber 2009). The authors should show such diagrams to provide a more complete and succinct view of the overall forecast performance of the case they selected.

(6) The case studies selected are cherry-picked given the confession that CTRL generally failed to provide reasonable forecast estimates of precipitation for both cases herein. For good measure, fairness and to better underscore the performance of the DA method, the authors should show the results for one case in which CTRL did not perform well and contrast it to one case where CTRL did preform reasonably well.

(7) The authors omit to mention that the degradation of the forecast at >= 3h is mainly due to saturation of the model solution by errors and biases within the initial / boundary conditions derived from large scale models or re-analysis datasets. This needs to be shown for both cases, especially given the unrealistically large (50 km) decorrelation length scale used for radar reflectivity factor.

(8) Title: Revise to include that: (i) primarily CG flashes are assimilated and (2) the model vehicle is RAMS.

Because these issues are collectively substantial and would require thorough rewriting of the manuscript in many places, I opted not to dwell on editorial comments for the time being. Additionally, the level of English remains, in my view, unacceptable for publication.

Figures:

Figures 5, 6, 8, 12a, 13a, 14a, 15a, 16a, 17a, 18a: The use of colored dots makes it very difficult to effectively compare the observations with those of the simulations: For consistency, either both sets of plots should show colored dots or shaded contours. For lightning, the authors should effectively show the gridded lightning data that were used to create the Qv or RH pseudo-observations.

Additional comments:

General comment: What is the main rationale for using a model that is marginally known by the community (RAMS) versus a more commonly used, battle tested, publicly available model such as WRF-ARW ? The authors not only seem to re-invent the wheel here but render any potential future work dedicated to reproducibility of the results - to the least - very challenging.

(1) Bottom, page 2: what are “conventional data” ? Why are radial velocity data not used ? Line 70: the main advantage of using 3DVAR vs 4DVAR, EnKF or hybrid methods lies in their already low computational burden. Thus, I do not agree with this justification. Also, variables are not “perturbed”; but adjusted by VAR methods.

(2) Pages 3 and 4: Please refer to Major Comments 1 and 3. Lines 105: Given that “Federico et al. (2017a) implemented the methodology of Fierro et al. (2012) …”, how come on line 112 “We use the method of Federico et al. (2017a) to assimilate lightning…” ? Please revise accordingly.

(3) Line 124: c.f. end of Major Comment 1.

(4) Line 240: RAMS used diagnostic relationships (vs explicit) to forecast lightning as it does not explicitly solves for the 3D electric field. Line 243: “Fourth”

(5) Line 290: Delete equation set as these are considered basic/common knowledge.

(6) Section 3.2, lines 300-312: Explicitly state and indicate that equation (2) is from Fierro et al. (2012, 2015) and not from Federico et al. Line 305: Please explain the rationales behind the choices of these constants: In particular, how are the forecast metrics affected for a 20% change in A, which has been shown to exhibit the most notable influence on the forced convection?

(7) Line 316-317: c.f. Major Comment 2.

(8) End of page 11: c.f. Major Comment 2

(9) Line 356: do the authors refer to the LFC or the LCL, (which may I add is an idea borrowed from Marchand and Fuelberg 2014 and Fierro et al. 2016). What is the top of the adjustment layer for lightning ? Please elaborate.

(10) Line 410 and elsewhere. This is similar to the results of Fierro et al. 2016. C.f. Major Comment 1. Please establish comparisons with previous works throughout the manuscript.

(11) Line 669: This statement is incorrect. The DE of ground based sensors levels off very rapidly with distance from land. This is where space-borne lightning detection systems such as the GLM or Feng Yun-4 can fill the gap.

(12) Lines 716-725: c.f. Major Comment 1.