Nonlinear Processes in Geophysics Discussions

Nonlin. Processes Geophys. Discuss., referee comment RC2
https://doi.org/10.5194/npg-2021-26-RC2, 2021

Comment on npg-2021-26

Anonymous Referee #2

Referee comment on "Direct Bayesian model reduction of smaller scale convective activity conditioned on large-scale dynamics" by Robert Polzin et al., Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2021-26-RC2, 2021

Review: Direct Bayesian model reduction of smaller scale convective activity

conditioned on large scale dynamics

General comments

The authors investigate the probabilistic impact of large scale atmospheric flow on small scale convective activity, using CAPE and vertical velocity as indicators, respectively. They apply a Direct Bayesian Model Reduction (DBMR) algorithm which was presented previously, to find so called latent states of the categorical input variable – the spatially averaged CAPE (averaged over $(500\text{km})^2$) and to estimate the probabilities of the categorical output – the number of small scale boxes with upward and downward vertical velocity conditioned on these latent states.

It becomes obvious that the authors intensively studied the topic. However, the manuscript suffers from many impressions and it is sometimes hard to keep track of the computational steps which are applied to data of different spatial and temporal scale.

The manuscript gives the impression, that it has been written in a hurry and I recommend to carefully review the manuscript internally and fix the numerous inaccuracies. Also, the readability should be improved. The captions of the figures are often not clear and the main text should be more to the point.

Unfortunately, the authors do not provide any assessment of the performance of the deduced model, so I find it hard to judge if the presented application of the DBMR yields useful results. Also, it would be nice if the authors could sketch, how the stochastic models deduced in the manuscript could be used in the future. It is mentioned, that there is in general the need for stochastic parametrization, but it is not made explicit, how this study concretely contributes to the issue. What has been achieved by deducing a stochastic model for updraft and downdraft (without any specification of intensity) on the small scale given a large scale value for CAPE?

For the presented study, the raw data must be assigned to categories. For what the authors term the ‘categorical input data’ the choice of categorization is comprehensible, even though the number of ‘causality boxes (n=10)’ is introduced ad hoc. For the ‘categorical output data’ the authors provide little on their choice of categories. Why are the three categories ‘updraft’, ‘downdraft’, ‘no draft’ important? And why does the intensity of the draft not matter in this study?

Regarding the data, as mentioned previously, it is sometimes hard to decipher which data was ultimately used. For example, Fig. 2 shows an area for which data is available, but seemingly all results presented in Sec.4 are solely based on the data from the North-West quadrant in Fig.2, if I understood correctly. Also, it unclear when hourly data plays a role and when if averages are taken before the computation of CAPE or afterwards. Or is CAPE given as a variable in the data set? Please see the specific comments to Sec. 3.

In Sec. 3.2.1 the authors propose two different ways to categorize the input data. However, in the remainder of the manuscript, apparently only one of these is considered – which one is not further specified.

The clearest section is Sec.2 which presents the model. However, this is not the work of this study, but rather a summary of Gerber and Horenko (2017). It remains unclear to me, why the method is considered to be Bayesian.

Is the fact that Γ*_{kj} □ {0, 1} (see l. 114) actually a constraint imposed on the algorithm or something that follows immediately from the definition of the model and the approximation of the log-likelihood given by Eq (9)? If it is a constraint, please elaborate on why this is chosen.

In this context, also the fact that the latent states are found by the algorithm itself, by assigning the Γ*_{kj} should be emphasized stronger and be revisited in Sec. 4.1.2.

Specific comments

l.5 The categorization is based on the conservation of total probability.

The meaning of the sentence is not very clear at this stage. What is ‘the categorization’? And, should the conservation of total probability not be a trivial fact?

l.9 Should it not be ‘reasearch on’ instead of ‘research of’?

l.15 Due to the geostrophic and hydrostatic equilibrium there is a scale separation induced by thermal stratification, gravity and rotation for scales above several kilometers (Klein, 2010).

Does this mean, that process of scale below several kilometers are separated from those of scale above several kilometers? Or can processes of scale above several kilometers be separated in different groups regarding to their spatial scale? This is a bit unclear.

Proposal: For processes of spatial scale above several kilometers, geostrophic and hydrostatic equilibria induce a temporal seperation of scales.

l.16 Medium-range forecasts are made up to 10 days in advance.

Medium range forecasts of what?

l.17 Predictions of convection further in advance cannot be deterministic and are highly uncertain because errors of the initial space of the smaller scales are growing.

What do you mean by ‘errors of the initial space? Do you mean errors in the estimate of the initial state?

The fact that errors grow is nothing special. Can you say something about the particularity of the error growth that apparently hinders forecasting?

l.19 A new perspective for improving General circulation models (GCMs) came from parameterizations.

Please clarify what is new. The use of parametrizations itself is not new.

l.21 Nowadays, many data-driven approaches are dwelling on stochastic parametrization methodologies involving the convective available potential energy (CAPE) as large scale driver for convection, e.g. in (Khouider et al., 2010; Dorrestijn et al., 2013a, b).

Is this citation in line with the citation style of the journal?

l.23 Their approaches need high computing capacities, but the costs to process large quantities of data can become a limiting factor.

It seems ‘but’ is the wrong conjunction here. Please replace by ‘and’ or clarify.

l.24 The statistical analysis of atmospheric dynamics simulations requires dimensionality reduction techniques which yield applicable reduced models.

This sounds as if ANY statistical analysis of such simulations would rely on dimensionality reduction. I am not certain, since this is not my field of expertise, but I doubt that this is true. Please formulate this statement more carefully.

l.29 The applicability of many approaches is based on the identification of reduced models defined on a small set of latent states.

This is a quite generic sentence. What are ‘many approaches’ here? Also, doesn’t this in parts duplicate the statement

‘The statistical analysis of atmospheric dynamics simulations requires dimensionality reduction techniques which yield applicable reduced models.’

from above (l.24)?

l.30 These methods derive aggregations of original variables based on a reduced approximation of the system in terms of relation matrices.

It is still not clear what methods this sentence refers to.

Both sentences, l.29 and l.30 receive meaning from the subsequent sentence. Not optimal, but okay.

l.42 Relation between the probability for large scale and smaller scales can be formulated categorically via a conditional probabilities and the conservation of the total probability.

I thought the large scale process will be treated as a given variable? Also, I think the word ‘relation’ requires an article here.

l.52 Our aim is to study and understand a stochastic relation between two variables X and Y that can take values from two finite sets.

In which way? Are the sets the same for both variables? Maybe, ‘that can take values from two corresponding finite sets each’?

Sec. 2.1 What about the stationarity of the processes? This should be a requirement for the method.

l.128 To apply DBMR, categorical processes for the in- and output have to be defined.

This sounds a little odd. Why would you ‘define’ processes? I thought the in- and output would be observational data? You do not specify the model of a categorical process, but rather you bin the observational data in order to make it categorical.

l.135 It is defined by Eq.(11) where θ_e is the pseudopotential temperature of the ascending air parcel, θ is the potential temperature of the surrounding air, and z_LFC is the so-called Level of Free Convection (LFC). The LFC is the height at which the rising air parcel becomes significantly warmer than its environment; Z_ET denotes the height, where the rising air parcel has the same temperature as its environment (ET stands for equal temperature). Thus, regarding its definition (11), CAPE becomes large if the temperature difference between the rising air and the environmental air is large, see (Bott, 2016, p. 431 ff).

The situation described here, is a little unclear. I assume z_LFZ < z_ET, correct? Further, we consider a air parcel, that is located at z_LFZ and that is about to rise, or one that is at height Z_ET? Or is this irrelevant? What is ‘significantly warmer than its environment’? I assume θ is a function of z? Otherwise, to which height does θ refer? What about θ_e? Is this a function of z? Maybe a little figure could be helpful.

l.151 To analyze the relation of large and small scale parameters, the COSMO-REA6 reanalysis data set is used (Bollmeyer et al., 2015).

This is very generic, you do not investigate the relation between small and large scale parameters in principle.

l.154 Since we focus on smaller scale convective events conditioned on large scale

dynamics in the atmosphere, we consider the summer months July and August in the years 1995 to 2015.

Why? Because these months feature most small scale convective events? Why do you neglect the rest of the time series?

l.157 This subdomain is bounded by the 45.2 â□¦ N – 54.7 â□¦ N , 5.8 â□¦ E – 15.3 â□¦ E and shown in Fig. 2.

I wonder if 'bounded by **the** [coordinates]' is a correct way to express an area.

Please, do not use the word 'subdomain' here, use 'domain' instead. The subdomains will be introduced at a later stage.

l.158 The Northwest coordinate is (5.8 â□¦ E; 54.7 â□¦ N) and the Southeast coordinate is (15.3 â□¦ E; 45.2 â□¦ N).

This is trivial and can be deleted.

Figure 2. Why are 500km x 500km quadrants so distinctly rectangular and not quadratic? Please check ‘domain’ wording.

l.162 The domain that covers Germany in Fig. 2 is divided into four 500 km × 500 km quadrants, where the spatial arithmetic mean of each of the quadrant is considered such that we obtain one CAPE value for each quadrant.

Do you average over the atmospheric variables provided by COSMO-REA6 and compute one value for CAPE subsequently? Or do you first compute a spatially dependent CAPE and average afterwards?

l.164 We separate and filter the data of CAPE and the vertical velocity in further subdomains in order to define the categorical in- and output.

What is meant by ‘separate and filter’ in this context?

After reading 3.2 the most likely interpretation seems to me, that for CAPE you use averages over the four quadrants, while for the vertical velocity you consider spatial averages over boxes of different sizes. But this is not clear from the text, especially since you write ‘we separate the data of CAPE and the vertical velocity in further subdomains’.

l.165 The corresponding sizes of the subdomains are summarized in Tab. 1.

Now the wording really becomes a mess. First, ‘subdomain’ is used for the entire region that covers Germany in l. 158. Second, it is used to refer what is called ‘quadrants’ in the text (l.162) and here it is used for even smaller ‘subdomains’?

l.166 For the analysis with DBMR, the northwest quadrant 1 over Holland in Fig. 2 is used.

Does this mean the rest of the data is not considered any further? Application of DBMR is the main purpose of this study as far as I understand. Then why do you introduce the entire data?

l.167 There is no influence of the Alps on smaller scale convective activity.

Does this refer to the 1 quadrant or to all quadrants?

l.197 There are exactly (m + 1)^2 ways to decompose m into the (ordered) sum of 3 nonnegative numbers.

I would not call this a decomposition. Rather, I would say, there are (m+1)^2 possibility to assign m observations to 3 different categories.

Why should the sum be ordered? If you consider an ordered assignment, the number should be larger.

Finally, I think you are missing a factor of ½. Please check.

l.198 In our probability-preserving algorithm the number of the occurring categories in the data are counted for the categorical observational input and output. The probability of a category is estimated by its occurrence frequency with respect to the total number of data points.

What ‘algorithm’ does this statement refer to? Apparently not the DBMR, since this algorithm relies on counting occurences. Does that mean that you compute Eq.(6) explicitly? If so, please explain why this is possible. In the beginning, you motivate the use of DBMR by saying, that this is often times computationally very costly.

Also, why does solving Eq.(6) require an ‘algorithm’?

l.207 We also evaluate the exact log-likelihood, as in (5). Fig. 3 shows the exact in blue and the relaxed log-likelihood in red, both for the reduced problem, i.e., the one with latent states.

Equation (5) to my understanding refers to the case without latent states. \Lambda in Eq.(5) must be replaced by \lambda * \Gamma to compute the likelihood of the reduced model.

l.209 The only parameter in the algorithmic procedure introduced above is the reduced process dimension K for the number of collective **causality boxes**.

You have not used the term ‘causality boxes’ before. Please introduce properly.

Also, what about ‘m’, the number of subdomains?

By the way, ‘m’ might not be the best choice for the number of subdomains, since in Sec. ‘m’ was introduced as the number of categories for the output variable.

Figure 3 – Caption: Again, Eq.(5) refers to the full model without latent states. Also, it is not clear what subdomains have been chosen to generate the results presented in the Figure.

Sec. 4.1.1

I do not understand, why the content of this section is presented under ‘results’ and not in the data pre- and postprocessing section. I think this would nicely fit into into 3.2.2, where the classification of w_i values is discussed in first place.

The histogram presented in Fig. 4 is not really a result of this work, but can legitimately be considered as part of preprocessing.

l.216 In the following, the pre- and postprocessing on DBMR with respect to the categorical in- and output will be discussed.

Why is this still ‘pre-and postprocessing’ if it is presented in the ‘Resuts’ section?

Also, what means ‘pre-and postprocessing **on** DBMR’?

l.218 All-day mean data serve as basis for determining the interval for vertical draft.

All-day mean data serve as **a** basis for determining the interval for vertical draft.

However, ‘the interval’ is an interval for the value of w_i that is classified as ‘no draft’. In the next sentence you refer to a ‘subclassification’. To me it is not quite clear, what exactly is being subclassified ‘by the interval’.

If possible, please reformulate in a more precise manner.

Fig.4 What exactly is shown here? In Sec. 3.1 you write that you use 12h averaged data, but the in caption it says hourly averaged data? Does the word ‘mean’ in ‘mean vertical velocities’ refer to a spatial mean over 125km? The data resolution is 6km! ‘vertical velocities for day and night’ contradicts ‘hourly averaged values’.

Please be more precise!

Red vertical lines represent a tube for **NO** vertical draft.

The summer months July and August in the years 1995 to 2015.

In captions it might be acceptable to write incomplete sentences, but this ‘statement’ is not convenient.

The sample size of the reanalysis data set sums up to S = 1302 (2 × 31 × 21).

Yes, but that is not what you are showing. Why do you actually consider hourly resolved data to define a classification for the 12h averaged data?

Also, you do not give any justification for the choice of the ‘no draft’ – interval. Please elaborate on the criteria that you apply. Is it a certain percentile in the data?

l.225 On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles of CAPE concerning the affiliated n categories (n = 10 divided by quantiles).

What is meant by ‘concerning the affiliated n categories (n = 10 divided by quantiles).’

l.227 In Fig. 5 one sees that the input categorization is similar in terms of value for day and night. The first latent state includes 5 (for day) and 4 (at night) CAPE categories with high values. This represents high CAPE values and is therefore referred to as “High”. Five (for day day) and 6 (at night) categories are affiliated to the second latent state, which is denoted with “Low”.

What criterion are these assignments based on? How are the latent states defined?

I think I understand this better having read the entire manuscript. However, the fact that first a manual categorization is performed and that the DBMR performs a second categorization might be the cause of some confusion here. This could be avoided if you

pointed out these two levels of categorization at an earlier stage.

Figure 5. Top: Boxplot of CAPE categories by 2 latent states for daily mean (left: day and right: night); Bottom: Affiliation of CAPE categories to the latent states; Data on large scale: 500 km × 500 km subdomains for northwest of Germany, hourly averaged CAPE; On mesoscale: 125 km × 125 km subdomains for vertical velocity. Time series: 21 years for July and August, S = 1302 (Length of time series)

I understand that in the top two panels, the boxplots show the 12h averaged CAPE data (spatially averaged over the Northwest quadrant of the COSMO-REA6 data) which is assigned to the latent states 'high' and 'low' on the left and right of each of the two top panels, respectively. The left panel shows the 'day' data and the right panel the 'night' data. However, this interpretation contradicts the fact, that the distributions shown for the 'high' and 'low' latent states do in fact overlap. Please clarify, what distributions are shown here?

This is not quite clear from the caption. Please use (a), (b) … to unambiguously mark the panels.

What role plays 'hourly averaged CAPE' in this plot?

What do the red crosses above the 75% quantile of the box plots mean? Are these

extreme events? Why are there none below the 25% quantile?

What does ‘On mesoscale: 125 km × 125 km subdomains for vertical velocity.’ mean in the context of this plot? No vertical velocity data is shown here.

l.233 The difference between the scales is small (375 km) with 500 km step size on large scale and 125 km step size on the smaller scale. The scale jump is of factor 4 on the basis of the small scale.

This is unclear. I understand that you somehow want to relate a (125km)^2 averaged vertical velocity data set with the (500km)² averaged CAPE data. But the (125km)² averaged vertical velocity data was not specifically introduced before? You are still at the stage, where you define the latent states for the categorical input data, why would you discuss specific choices for the output data resolution, here?

Please clarify at the beginning of the section, that a) you now move to the application of the DBMR algorithm and b) that this application requires a choice of the scale of the categorical output or in other words, a choice of the spatial average taken on the vertical velocity data. Please clarify that for the first DBMR application the spatial scale of the output is set to 125km.

Also, it would help to emphasize, that the choice of the spatial scale for the categorical output will influence the latent states identified by the DBMR.

l.245 In Fig. 6, K bivariate histograms are shown for day and night respectively.

This is a specific case where apparently K=2, so please indicate.

Figure 6: Please use a discretized colorbar. Also check if other colorbars facilitate a better distinction between the many low values. Please use (a), (b) ,… to identify panels

To which quadrant of the COSMO-REA6 data do the results refer to?

l. 256 In the top left panel (Z 1 , day) of Fig. 6 probability adds up for numbers of up- or downdrafts higher than 10 to 81%. In the top right panel, probability accumulates at small numbers of boxes with downdraft.

Judging from an optical assessment there should be at least as much probability weight allocated to states where both, the number of downdraft cells and the number of updraft cells are below 10. This contradicts the above statement. Please check.

l.257 In the top right panel, probability accumulates at small numbers of boxes with downdraft.

I cannot see any of the states with pronounced downdraft having high probability! Much of the probability mass is allocated to states with no downdraft, and little updraft.

l.262 In the bottom left panel (Z 2 , night) the probability is accumulated to 82% for the number of updrafts between 0 and 4.

That’s the bottom right panel.

Figure 7: What is the difference between Fig.5 and Fig.7 ? If it is, that you use CAPE values averaged spatially averaged over $(125km)^2$ in fig.5 and $(15km)^2$ in fig 7, then this is not clear and also contradicts the statements in sec. 3:

The domain that covers Germany in Fig. 2 is divided into four 500 km × 500 km quadrants, where the spatial arithmetic mean

of each of the quadrant is considered such that we obtain one CAPE value for each quadrant.

We use the average of the 500 km × 500 km

quadrants, considering CAPE as the large scale atmospheric driver.

l.270 KDE is a non-parametric way to estimate the probability density function of a random variable.

A KDE requires the choice of a bandwidth and is therefore not non-parametric.

l.283 Affiliations without gaps lead to a separation of the latent states. “No gaps” means that affiliations are interrelated and not interrupted in the middle plots of Fig. A1 and Fig. A3. The affiliations have no gaps for day and night.

Of course the latent states should cover all categorical input states and should have ‘no gaps’. Or does ‘no gaps’ mean that there is an overlap of spatially smaller scale CAPE values which are assigned to the different latent states based on the large scale CAPE average?

l.343 The representation of probabilities of numbers of updrafts and downdrafts conditioned on the latent states for the convective scale in Fig. 8 correspond in their distributions to the results on mesoscale in Fig. 6.

'for the convective scale' should refer to 'probabilities of numbers of updrafts and downdraft', right? In the current version of the sentence, 'for the convective scale' seemingly refers to 'the latent state'.

l.355 The stochastic method is tested in a meteorological application towards a model reduction to latent states of smaller scale convective activity conditioned on large scale atmopsheric flow.

'latent states of smaller scale convective activity' suggests that the latent states are introduce for the vertical velocity, however, they have been introduced for the caterogical input, which is the large scale atmospheric flow, that is CAPE.

Technical comments

l.26 Since its introduction by Lorenz (1956), EOF analysis—known as principal component analysis (PCA) or proper orthogonal decomposition (POD)—has become an important statistical tool in atmosphere science.

- also known … -

l.37 The latter approach does not require a distributional assumption but works instead with a discretized state vector.

Why ‘the latter’?

l.43 Various energetic variable are applicable on large scale.

variables is missing an s.

l.48 In Sect. 4 the results are discussed related to atmospheric dynamics.

… the results are discussed **and** related to

or the results are discussed with regard to atmospheric dynamics.

l.114 Moreover, the method yields Γ*_{kj} ☐ {0, 1}, i.e., the original input categories are assigned to the reduced system's (latent) categories in a deterministic fashion (no "fuzzyness" in the affiliations).

Why is the Γ*_{kj} marked with an *?

l.120 This manifests in the variance of the estimated parameter λ ☐ ik , which shows a K/n-times smaller uncertainty than Λ ij…

I am not sure the use of 'manifest' is correct here.

l.155 The sample size of the reanalysis data set used in Sect. 2 sums up to S = 1302 (2 × 31 × 21)

Sect. 2? Wrong reference?

l.168 According to the meteorological data in Sect. 3.1…

According to the meteorological data **described** in Sect. 3.1…

l.169 CAPE plays the role of an input variable X in Sect. 2

CAPE plays the role of an input variable X **as defined** in Sect. 2

l.175 With this type of classification, extreme weather events tend to be in a separate category.

tend to **be in separate categories**.

l.175 These are not Gaussian distributed.

Grammatically the reference of ‘these’ is not clear. Could refer to ‘categories’ from the previous sentence. I assume you meant to refer to ‘extreme weather events’. Please clarify and add ‘in terms of CAPE’.

l.193 Let Y_i (t) be the discretized vertical velocities at time t with $1 \leq i \leq m$ numbering the grid boxes on the corresponding scale, see Tab. ??.

Probably, it is better to not use m as an upper bound for i, since m was already use for the number of observations of Y in time. Here, m refers to the spatial number of observations that are being made simultaneously, if I understand correctly.

Please correct the reference.

l.206 for every fixed number K of latent state.

for every fixed number K of latent state**s**.

l.206 For the respective latent state...

For each K.

‘The latent state’ would refer to an individual state.

l.220 In Fig. 4, the histogram of mean vertical velocities for a resolution of 125km with the interval is shown.

In Fig. 4, the histogram of mean vertical velocities for a resolution of 125km is shown together with the interval that defines the ‘no draft’ category.

l.229 Five (for day day) and 6 (at night) categories

double day

l.244 In order to visualize the probabilities of the small scale conditioned on the latent states…

In order to visualize the probabilities of the small scale variable conditioned on the latent states of the large scale variable…

l.244 the entries of the λÌ□

why does \lambda carry a hat?

Please also note the supplement to this comment:
https://npg.copernicus.org/preprints/npg-2021-26/npg-2021-26-RC2-supplement.pdf