Direct Bayesian model reduction of smaller scale convective activity conditioned on large scale dynamics

Robert Polzin[1], Annette Müller[2], Henning Rust[2], Peter Névir[2], and Péter Koltai[1]

[1]Institute of Mathematics, Free University Berlin, Germany
[2]Institute of Meteorology, Free University Berlin, Germany

Correspondence: Robert M. Polzin (robert.polzin@fu-berlin.de)

Abstract.

We pursue a simplified stochastic representation of smaller scale convective activity conditioned on large scale dynamics in the atmosphere. For identifying a Bayesian model describing the relation of different scales we use a probabilistic approach by Gerber and Horenko (2017) called *Direct Bayesian Model Reduction* (DBMR). **R2SC1:** This is a Bayesian relation model between categorical processes (discrete states), formulated via the conditional probabilities. The convective available potential energy (CAPE) is applied as large scale flow variable combined with a subgrid smaller scale time series for the vertical velocity. We found a probabilistic relation of CAPE and vertical up- and downdraft for day and night. This strategy is part of a development process for parametrizations in models of atmospheric dynamics representing the effective influence of unresolved vertical motion on the large scale flows. **R2SC2:** The direct probabilistic approach provides a basis for further research on smaller scale convective activity conditioned on other possible large scale drivers.

1 Introduction

Complex dynamical processes involving scaling cascades are omnipresent in natural science. Such processes feature different characteristic scales. The smallest and largest scales are far apart and much of the scale range is involved by scale interactions. Dynamics in the atmosphere take place across a large range of time- and length scales, from micro-seconds to months and lengths from 10^{-5} to 10^{6} m. **R2SC3:** For processes of spatial scale above several kilometers, geostrophic and hydrostatic equilibria induce a spatial-temporal separation of scales; see Klein (2010). Thunderstorms last a few tens of minutes for example, whereas hurricanes may last for days. **R2SC4:** Medium-range weather forecasts are made up to 10 days in advance. **R2SC5:** Predictions of convection further in advance cannot be deterministic and are highly uncertain because errors of the variable on small scale at the initial state are growing.

R2SC6: A new perspective for for weather and climate models came from stochastic parameterizations that represent the small scale effects of convection on the large-scale dynamics; see Berner et al. (2017); Franzke et al. (2015). For instance, Gottwald et al Gottwald et al. (2016) parametrize in the tropics convective area fraction conditioned on large scale vertical velocity. Also, many data-driven approaches consider stochastic parametrization methodologies involving the convective available potential energy (CAPE) as large scale driver for convection; see Khouider et al. (2010); Dorrestijn et al. (2013a, b).

R2SC8: Their approaches need high computing capacities and the costs to process large quantities of data can become a limiting factor. **R2SC9:** Some statistical analyses of atmospheric dynamics simulations requires dimensionality reduction techniques which yield applicable reduced models; e.g. Horenko (2008). One way is the Empirical orthogonal function (EOF) analysis which is a tool for data compression and dimensionality reduction used in meteorology. **R2TC1:** Since its introduction by Lorenz (1956), EOF analysis—also known as principal component analysis (PCA) or proper orthogonal decomposition (POD)—has become an important statistical tool in atmosphere science. For example in Horenko et al. (2008) different sets of EOFs are used for a reduced representation of meteorological data. **R2SC11:** Other examples for reduced approximation in terms of relation matrices are covariance matrices (Schölkopf et al., 1997; Jolliffe, 2003), partial autocorrelation matrices of autoregressive processes (Schmid, 2010), Gaussian distance kernel matrices (Donoho and Grimes, 2003; Coifman et al., 2005), Laplacian matrices as in the case of spectral clustering methods for graphs (Von Luxburg, 2007), adjacency matrices in community identification methods for networks (Zhao et al., 2012); see Gerber and Horenko (2017). A recent algorithmic framework called Direct Bayesian Model Reduction (DBMR) provides a computationally scalable probability-preserving identification of reduced models and latent states directly from the data; see Gerber and Horenko (2017); Gerber et al. (2018). The method constructs a directly low-rank transition matrix, reducing numerical effort and estimation error due to finite data. **R2TC2:** The approach does not require a distributional assumption but works instead with a discretized state vector. Our aim is the development of a model combining the deterministic large scale atmospheric flow with a conceptual stochastic description of small scale convection. Towards this goal, we develop a conceptual categorical description for smaller scale vertical velocity, which is linked to a large scale flow variable. The probabilistic description is proposed using DBMR. **R2SC12:** This Bayesian relation model between large scale and smaller scales can be formulated categorically via a conditional probabilities in the law of total probability. **R2TC3:** Various energetic variables are applicable on large scale. Other potential large scale variables driving the smaller scale stochastics besides CAPE are the Dynamic State Index (DSI) in Müller et al. (2020) and Müller and Névir (2019), available moisture, or vertical wind shear. The DSI is a scalar diagnostic field that quantifies local deviations from a steady and adiabatic wind solution and thus indicates non-stationarity as well as diabaticity.

The paper is structured as follows: In Sect. 2 the mathematical methodology of DBMR is presented. Afterwards, in Sect. 3 the set-up for a reduced model in the atmosphere is described. **R2TC4:** In Sect. 4 the results are discussed with regard to atmospheric dynamics. Finally, in the conclusion the results and future work towards the direct Bayesian model reduction of smaller scale convective activity conditioned on large scale dynamics are formulated.

2 Mathematical methodology

Our aim is to study and understand a stochastic relation between two variables X and Y that can take values from two finite sets. **R2SC13:** The sets of both variables can be different, as we discuss in our meteorological applications. We assume that the probabilistic dependence of Y on X is time-independent. Whether X_t and Y_t as t-parametrized stochastic processes are themselves stationary does not play a role here. The categorical random variables X and Y will later on encode quantitative information of the atmosphere on different spatial scales. We will review a novel computational framework for the estima-

tion of a reduced (low-rank) Bayesian model from data. This method is called Direct Bayesian Model Reduction (DBMR). Direct refers to a directly low-rank estimation which is useful for the identification of reduced models, yielding thereby an advantageous estimation error, especially if data is not abundant; see Gerber and Horenko (2017).

2.1 R1C1, R2GC7: Full Bayesian model formulation

We are interested in modeling the probabilistic relationship of two potentially random quantities, X and Y. For us, it will be only relevant that Y is a random function of X—randomness of X itself is irrelevant. Since the observations typically arise as time series, we consider X and Y as processes $X(t)$ and $Y(t)$ with time t, however t can denote any parameter ordering the realizations of the process. We will consider the case where X and Y can only attain a finite number of values, such that we call the processes discrete-state or categorical. Say, $Y(t)$ is taking one of the possible values from m categories $\{y_1, y_2, ..., y_m\}$ and $X(t)$ from the n categories $\{x_1, x_2, ..., x_n\}$. The central quantity of interest describing the relationship of X and Y is the $m \times n$ matrix of conditional probabilities, also called transition matrix,

$$\Lambda = \begin{pmatrix} \mathbb{P}[Y = y_1 \,|\, X = x_1] & \cdots & \mathbb{P}[Y = y_1 \,|\, X = x_n] \\ \vdots & \ddots & \vdots \\ \mathbb{P}[Y = y_m \,|\, X = x_1] & \cdots & \mathbb{P}[Y = y_m \,|\, X = x_n] \end{pmatrix}. \tag{1}$$

Note that Λ is a column-stochastic matrix. In practical studies, when the Λ_{ij} are estimated from the available observations of X and Y one needs to guarantee that the data is acceptably randomised; see Holland (1986). We will assume that

$$\text{Law}\left[Y(t) \mid X(1), X(2), \ldots\right] = \text{Law}\left[Y(t) \mid X(t)\right], \tag{2}$$

i.e., given the input $X(t)$, the distribution of the output $Y(t)$ is independent of the other inputs.

2.2 R1C1, R2GC7: Maximum likelihood approach

Typically, the transition matrix Λ is not directly available and can only be estimated from observed data. Let S be the number of observation pairs for the categorical processes X and Y, such that the following observational data is available:

$$\boldsymbol{XY} = \{X(1), X(2), ..., X(S), Y(1), Y(2), ..., Y(S)\}, \tag{3}$$

where $X(t) \in \{x_1, \ldots, x_n\}$ and $Y(t) \in \{y_1, \ldots, y_m\}$, as above. Given $\boldsymbol{XY}$, it is reasonable to search for the Λ for which the total probability of obtaining the particular sequences of observations (3) is maximized. By the independence assumption (2), the *likelihood* of a matrix Λ of conditional probabilities—i.e., the probability of observing the data if the conditional probabilities were given by Λ—is given by

$$\mathbb{P}[\boldsymbol{XY} \,|\, \Lambda] \;\propto\; \prod_{i=1}^{m} \prod_{j=1}^{n} \underbrace{\mathbb{P}[Y = y_i \mid X = x_j]}_{=\Lambda_{ij}}^{N_{ij}}, \tag{4}$$

where N_{ij} is the total number of instances in the data when $\big(X(t), Y(t)\big) = (x_j, y_i)$. The optimum can be more easily computed if one considers the *log-likelihood* $\log(\mathbb{P}[\boldsymbol{XY} \mid \Lambda]) = \sum_{i=1}^{m}\sum_{j=1}^{n} N_{ij} \log \Lambda_{ij}$, with which we arrive at the maximum likelihood problem

$$\Lambda^* = \arg\max_{\Lambda} \left\{ \sum_{i=1}^{m}\sum_{j=1}^{n} N_{ij} \log \Lambda_{ij} \right\}, \text{ such that } \Lambda_{ij} \geq 0, \sum_{i=1}^{m} \Lambda_{ij} = 1. \tag{5}$$

The optimal solution of that constrained optimisation problem can be determined analytically (Gerber and Horenko, 2017), resulting in the empirical frequency estimator:

$$\Lambda^*_{ij} = \frac{N_{ij}}{\sum_j^n N_{ij}}. \tag{6}$$

Since we merely have a finite amount of observation at hand, it is essential to be aware of the uncertainty of the statistical estimate (6). While we refer the reader interested in exact bounds to Gerber and Horenko (2017, Supplement, Eq. (14)), an intuition can be gained as follows. To estimate each Λ_{ij} to a sufficient (statistical) accuracy, the transitions N_{ij} should be, on average, numerous. As there are nm parameters in Λ to estimate, this asks for the sample size S to be reasonably large as compared to nm. In practice, this can be problematic if n and m are large. Thus, next we will discuss a modification of the above method that can mitigate this problem.

2.3 R2GC2: Model reduction to latent states

In numerous situations the apparent complexity of our observations is an artefact of our measurement procedure, and there are low-dimensional features that govern the process at hand. Thus, even if we would be able to find a full matrix Λ of conditional probabilities, the ultimate goal would be to reduce this through such low-dimensional features.

The following approach, proposed by Gerber and Horenko (2017), achieves both estimation and reduction in one step. We *assume* that the output depends on the input through a *latent variable* Z, which can merely take a small number $K \ll \min\{n, m\}$ of different states $\{z_1, \ldots, z_K\}$. In terms of probabilistic influences, we assume the structure

$$X \xrightarrow{\Gamma} Z \xrightarrow{\lambda} Y, \tag{7}$$

where λ, Γ are matrices of conditional probabilities,

$$\Gamma_{kj} = \mathbb{P}\big[Z = z_k \,|\, X = x_j\big], \qquad \lambda_{ik} = \mathbb{P}\big[Y = y_i \,|\, Z = z_k\big]. \tag{8}$$

We also assume conditional independence of Y on X given Z, that is, the input-output conditional probability matrix Λ satisfies $\Lambda = \lambda\Gamma$. Note that we can interpret Γ_{kj} as an *affiliation* of input category x_j to the latent state z_k, see Fig. 1.

The task is now to determine the pair of column-stochastic matrices (λ, Γ) from the observation data $\boldsymbol{XY}$, as given in (3). Again, we wish to solve the problem with a maximum-likelihood approach, which would require solving (5) with replacing Λ_{ij} by $(\lambda\Gamma)_{ij}$ and the constraints by requiring λ and Γ being stochastic matrices. This, however, is a computationally hard

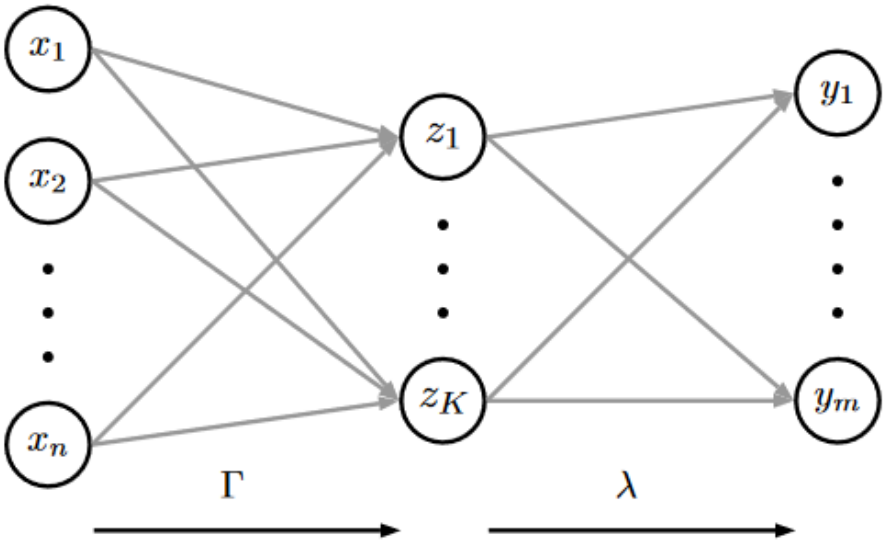

Figure 1. Introduction of intermediate latent states in DBMR for efficient and scalable estimation of Λ

optimization problem, which in Gerber and Horenko (2017) relax to

$$(\lambda^*, \Gamma^*) = \arg\max_{\lambda,\Gamma} \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{K} N_{ij} \Gamma_{kj} \log\{\lambda_{ik}\} \tag{9}$$

subject to

$$\lambda_{ik} \geq 0, \sum_{i=1}^{m} \lambda_{ik} = 1, \qquad \Gamma_{kj} \geq 0, \sum_{k=1}^{K} \Gamma_{kj} = 1. \tag{10}$$

While (9) produces suboptimal estimates, its advantage comes from the fact that it is concave in both variables λ and Γ, respectively, allowing for a very simple alternating maximization as optimization procedure; see Gerber and Horenko (2017). The resulting algorithm is DBMR. Moreover, the method yields $\Gamma^*_{kj} \in \{0, 1\}$, i.e., the original input categories are assigned to the reduced system's (latent) categories in a deterministic fashion (no "fuzzyness" in the affiliations). **R2GC8:** The binary nature is a property of the optimal solution; see Gerber and Horenko (2017). Of course, the number K of latent states is not known in advance, and has to be chosen judiciously by compromising between "expressiveness" (the likelihood of the model, i.e., the optimal value in (9)) and "effort" (the number of total parameters to be estimated and their statistical error). This can be done comparing multiple DBMR runs with different K.

The obtained models are also less subject to overfitting issues and are more advantageous in terms of the model quality measures (Gerber and Horenko, 2017; Gerber et al., 2018). **R2TC6:** This is expressed in the variance of the estimated parameter λ^*_{ik}, which shows a K/n-times smaller uncertainty than Λ_{ij}; see Gerber and Horenko (2017, Theorem and eqn. [7]). Again, intuitively this advantage of DBMR over the full model (6) can be seen by noting that from the same amount of data DBMR only needs to estimate $k(n+m)$ parameters, while the full model nm parameters.

Let us emphasize that additionally to all the computational advantages of DBMR that allow it to work with large data sets, its conceptual strength is that it combines model estimation and model reduction in one step. The latent states often have a physical meaning—a property that we shall focus on in our application.

3 R2GC2: Meteorological data processing

R2SC14: To apply DBMR, a quantization of the input and output processes into categories has to be performed. First, we discuss the choice of meteorological variables and scales in view of the categorical processes. As input we use a variable related to large scale atmospheric flow: *Convective Available Potential Energy* (CAPE), a measure for the energy an air parcel would gain if lifted to a specific height in the atmosphere.

3.1 R2GC2: Scales, variables and data preprocessing

CAPE can be seen as a measure for atmospheric stability, first suggested by Weisman and Klemp (1982). It is defined by

$$CAPE = g \int_{z_{\mathrm{LFC}}}^{z_{\mathrm{ET}}} \frac{\theta_e - \theta}{\theta}\, dz\,, \tag{11}$$

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). **R2SC15:** For positive CAPE, this difference must be positive. CAPE is determined by the layer thickness between the starting and ending points in space (height) and by the integrand in (11). Boundary conditions can vary. θ can be a function of z, it depends on the difference between the heights and the potential temperature. As an integral, CAPE is a global variable that we consider as representative variable on the larger scale. To capture convective activity, characterized by strong up- and downdrafts, on the smaller scale, we regard the vertical velocity. Parcel theory predicts

$$\mathrm{CAPE} \sim \textbf{R1C2:} \frac{w_{max}^2}{2}, \tag{12}$$

where w_{max} is the maximum vertical motion in the dimension m/s expected from the release of CAPE in the dimension J/kg; see Dutton (1976). The relation in Eq. (12) is a kinetic description of a potential which does not have to be released to vertical updraft. Moncrieff and Miller (1976) were the first to use the term CAPE. The USAF Air Weather Service (which changed its name to the Air Force Weather Agency in 1997) simply called it positive area; see Blanchard (1998). Fritsch and Chappell (1980) called it potential buoyant energy (PBE), while variations of this include +BE and net positive buoyant energy. Despite the abundance of names, it now appears that CAPE is the de facto standard terminology. In Kirkpatrick et al. (2009) over 200 convective storm simulations are analyzed to examine the variability in storm vertical velocity and updraft area characteristics as a function of basic environmental parameter CAPE.

R2SC16: For our studies, the COSMO-REA6 reanalysis data set is used; see Bollmeyer et al. (2015). This reanalysis is based on the non-hydrostatic numerical weather prediction model COSMO (*COnsortium for Small scale MOdelling*) by the German Meteorological Service (Deutscher Wetterdienst, DWD) using a continuous nudging scheme. It has a horizontal resolution of 6 km and 40 vertical layers; see Bollmeyer et al. (2015). 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. **R2SC17:** The summer months are predestinated for convective events. The months from May to August are possible. We only worked with two months in order not to have too much data. **R2GC1, R2SC20:** For our analysis the raw data are hourly REA6 data. We first computed CAPE as REA6 variable and then averaged for the respective spatial scale to 12 hours means. **R2TC7:** The sample size of the reanalysis data set used in Sect. 2.2 sums up to $S = 1302$ $(2 \times 31 \times 21)$. Moreover, REA6 data is available for Germany. **R2GC5,R2SC23:** In order to focus our method we started at the top left with the first quadrant; see Fig. 2. Here we expect the relatively flat surface in the north of Germany to be more homogeneous and different from the pre-alpine southern regions with forced uplifting. **R2SC18:** The top left quadrant is bounded by the $[45.2°N$ to $54.7°N$, $5.8°E$ to $15.3°E]$ and shown in Fig. 2. The Northwest coordinate is $[5.8°E;\ 54.7°N]$ and the Southeast coordinate is $[15.3°E;\ 45.2°N]$. As vertical layer the 600 hPa surface is considered, because here the latent heat release takes place and the vertical velocity reaches its maximum; see Müller et al. (2020).

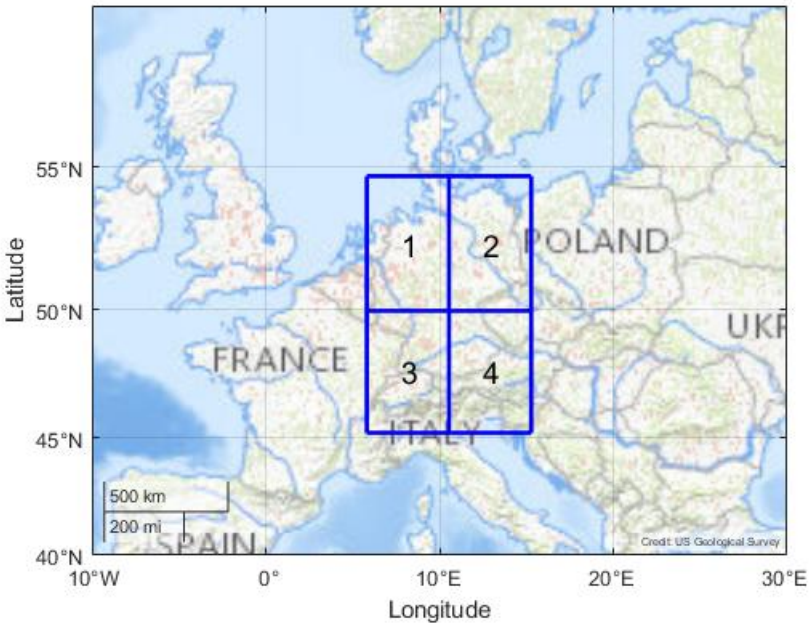

Figure 2. REA6 domain that covers Germany consisting of grid boxes 1 to 4; Grid box 1 is applied on the large scale for DBMR and is of approximately $500\,\mathrm{km} \times 500\,\mathrm{km}$. Image credit of the map: US Geological Survey (USGS).

3.1.1 Filtering CAPE and vertical velocity

R2SC22: We used the term 'domain' for the total region we considered in Fig. 2, 'quadrant' for the respective quarters on large scale of it, and 'grid boxes' for the other partitions of the domain on smaller scales. For the large scale, the domain that covers Germany in Fig. 2 is divided into four $500\,\mathrm{km} \times 500\,\mathrm{km}$ quadrants. For smaller scales, the quadrants are refined in a_m grid boxes of different sizes; see Tab. 1.

R2SC28: #Grid boxes a_m	Edge length
$1^2 = 1$	1000 km
$2^2 = 4$	500 km
$4^2 = 16$	250 km
$6^2 = 36$	167 km
$8^2 = 64$	125 km
$10^2 = 100$	100 km
$12^2 = 144$	83 km
$16^2 = 256$	63 km
$32^2 = 1024$	31 km
$64^2 = 4096$	15 km

Table 1. Number of grid boxes a_m and edge length for box discretization of the atmosphere from large synoptic scale (1000 km) across intermediate scales up to meso-gamma scale with convective activity ($2 - 20$ km)

3.2 R2GC2: CAPE and vertical velocity as in- and output data

R2TC8: According to the meteorological data described in Sect. 3.1, we will set up applicable categories for in- and output. **R2TC9:** CAPE plays the role of an *input* variable X as defined in Sect. 2, describing the potential for convection. **R2SC23:** The spatial arithmetic mean of each of the $500\,\mathrm{km} \times 500\,\mathrm{km}$ quadrants such that we obtain one CAPE value for each quadrant as the large scale atmospheric driver. With energy units, CAPE has a non-negative range of values. The model's *output* variable Y is vertical velocity obtained on a smaller scale. Here, Y can take positive and negative values for updrafts and downdrafts, respectively. We average over $250\,\mathrm{km} \times 250\,\mathrm{km}$ to $15\,\mathrm{km} \times 15\,\mathrm{km}$ according to Tab. **R1C5:** 1.

3.2.1 Categorical input

R2GC6: We consider the range of values for CAPE (X) and generate n categories by $\{x \in X_i \mid b_{i-1} \leq x < b_i\}$. For the category boundaries b_i, we consider the following spaced option in probability using empirical $1/n$-quantiles as category boundaries b_i. This categorization has the advantage of (almost) equally populated categories. The resulting n categories are denoted by integers $1, ..., n$. **R2GC4:** The chosen categorization depends on the amount of available data. We varied input

numbers and choose $n = 10$ by subjective physical plausibility. In Sect. 3.1 we set up the meteorological data with a size of the observational data $S = 1302$. That means we have about 130 data points in each CAPE category.

3.2.2 Categorical output

We map vertical velocities ω_i at grid box i on a variable $Y_i \in \{1,2,3\}$ as

- **updraft** for $Y_i = 1$, if $\omega_i \geq a_2$,
- **no draft** for $Y_i = 2$, if $a_1 \leq \omega_i < a_2$,
- **downdraft** for $Y_i = 3$, if $\omega_i < a_1$,

$(a_1, a_2) \in \mathbb{R}^2$ define a potentially asymmetric interval around zero vertical velocity which we consider as neutral, with $a_1 < 0$ and $a_2 > 0$. **R2GC4:** The predefinition of 'updraft', 'downdraft', 'no draft' determines whether there is convection and, if so, how it is directed (upwards or possibly downwards). The choice of (a_1, a_2) depends on the scale of the box where Y is averaged over. In Sect. 4.1 the choice of the interval for our analysis is described. Once this discretization is made, the final output categories needed can be set up. **R2TC12:** Let $Y_i(t)$ be the discretized vertical velocities at time t with $1 \leq i \leq a_m$ numbering the grid boxes on the corresponding scale, see Tab. 1. We define the following categorical process

$$\hat{Y}(t) = (\#\{Y_i(t) = 1\}, \#\{Y_i(t) = 2\}, \#\{Y_i(t) = 3\}) \in \mathbb{N}^3, \tag{13}$$

with $\#\{Y_i = k\}$ being the number of grid boxes with vertical velocity mapped onto $k \in \{1,2,3\}$. **R2SC25:** Note that $\sum_{k=1}^{3} \#\{Y_i = k\} = m$, the number of grid boxes. There are exactly $(a_m + 1)^2$ ways to decompose a_m into the (ordered) sum of 3 nonnegative numbers, thus the number of actually occurring categories $n_{\hat{Y}} \leq (a_m + 1)^2$. **R2SC26:** In DBMR, the numbers of actually occurring categories are counted. These numbers have impact on the probability distribution of the categories for input and output. We try to conclude down- and updraft behavior from the $\hat{Y}(t)$, i.e. the distribution of up- and downdrafts. Note that we have in this experiment no information on the (spatial) structure, as the category in (13) is a triple of numbers for counts of down-, updraft and low vertical velocity.

3.2.3 R2GC2: Interval for vertical draft

R2SC32: 12h mean data for day and night serve as basis for determining the interval for vertical draft which was chosen symmetrically with interval limits $a_1 = -0.0048$m/s and $a_2 = 0.0048$m/s. **R2TC15:** In Fig. 3, the histogram of mean vertical velocities for a resolution of 125km is shown together with the interval that defines the 'no draft' category. For the application of DBMR the data for day and night are split up and will be applied separately with the same interval for vertical velocity.

3.3 R2GC2: Reliability and assessment of performance

The model reduction is a consequence of using the affiliation matrix Γ, which assigns the n large scale categories to $K < n$ latent states. In the frame of DBMR we optimize a relaxed log-likelihood, cf. (9). **R1TC13:** We ran DBMR 100 times (with

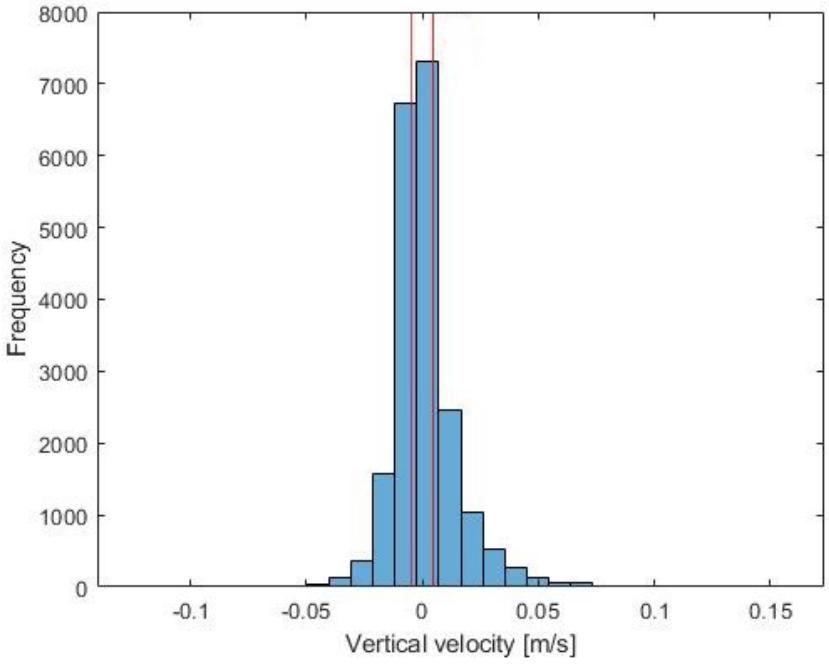

Figure 3. **R2SC30:** Histogram of spatial mean vertical velocities for day and night for a resolution of 125 km (64 grid boxes on small scale) is presented; The red vertical lines show the interval for vertical velocity which was selected on the basis of the 75th-percentile of the data set. The sample size of the 12h mean data set for day and night sums up to $2S = 2604$.

random initializations) for every fixed number K of latent states. **R1TC14:** For each K, the run with the maximum log-likelihood is presented. **R2SC27:** We also evaluate the exact log-likelihood, as in Eq. (5) which refers to the case without latent states. Λ in Eq. (5) was replaced by $\lambda\Gamma$ to compute the likelihood of the reduced model. Fig. 4 shows the exact in blue and the relaxed log-likelihood in red, both for the reduced problem, i.e., the one with latent states. **R2SC28:** The only parameter in the algorithmic procedure introduced above is the reduced process dimension K for the number of latent states. It can be chosen by comparing results for different K and selecting the best reduced model according to one of the standard model selection criteria (Cross-validation with a performance criterion, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) or L curve approach). For an attempt for model selection, the largest increase in log-likelihood can be found by increasing $K = 2$ to $K = 3$, for $K = 6$, the maximum value has been reached. Note that as $n = 10$, choosing $K = 10$ presents no model reduction. **R1C4:** In view of uncertainties and biases for parameterization vertical velocity can be hard to measure and is likely to be biased in reanalyses. We work with discretize vertical velocity and thus with a less precise variable. This makes the problem of uncertainty and bias less relevant but is definitively not a relief. In a stochastic model for the updraft which is to be developed, one can think of including an additional parameter as factor to the vertical velocity to allow for a tuning with respect to the effect generated by the modelled updraft. **R2GC3:** Although quantification of model performance is

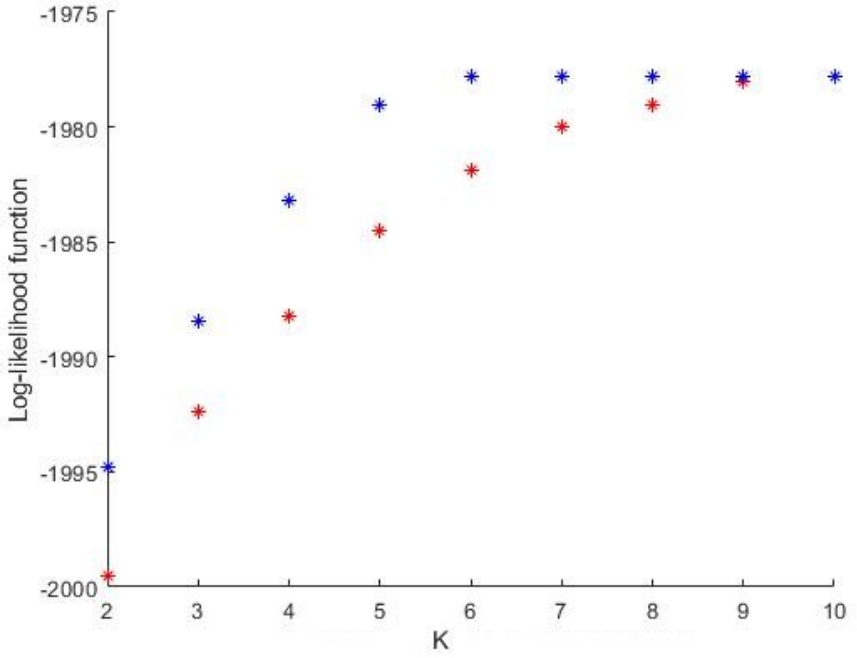

Figure 4. **R2SC28:** Exact log-likelihood value as in (5) (blue) and relaxed log-likelihood value as in (9) (red) of the reduced Bayesian model estimated by DBMR with K latent states for day for a resolution of 125 km (64 grid boxes on small scale).

possible here using, e.g., a cross validation study given an adequate score of interest, it is probably not very helpful at this stage. We consider our study rather as a proof-of-concept ideally preparing grounds for a stochastic model for vertical movement to be inserted into a circulation model. Usefulness should be evaluated then in terms of circulation model simulations. Further work is required to give the latent states a meteorological meaning in the sense of circulation weather types, regarding all seasons separately.

4 R2GC2: Reduced model for convective activity

4.1 R2GC2: Dynamics separated by day and night

4.1.1 Affiliation to latent states

R2SC36: In Fig. 5 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. **R2SC34:** 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 the 10 CAPE

categories. **R2SC36:** The 25th percentiles of the distributions shown for the 'High' latent states overlap the 75th percentiles of distributions shown for the 'Low' latent states. The first latent state includes 5 (for day) and 4 (at night) CAPE categories with 'High' values. **R2TC16:** Five (for day) and 6 (at night) categories are affiliated to the second latent state, which is denoted with 'Low'. **R2GC9, R2SC35:** The structure of the affiliation Γ^* in (9) is assigned by minimizing information entropy (the likelihood bound). The question of how the latent states are defined and what criterion are the affiliations based on, we emphasize that the latent states are found by the algorithm itself. During daytime the categories reach values up to 386 J/kg, whereas at night the values have a range of 343 J/kg due to less convective activity; see boxplots in Fig. 5. The affiliations to the latent states have no gaps for day and night. That means the latent states are separate from each other. **R2SC37:** The choice of the spatial scale for the categorical output will influence the latent states identified by the DBMR. 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. The results for three latent states are shown in Appendix A. There is a third latent state which represents 'Mean' CAPE categories; see Fig. A1. In the following, the output of the DBMR is discussed.

4.1.2 Distributions conditioned on latent states

We discuss probability distributions conditioned on the resulting latent states introduced in Sect. 2.3 in two ways:

- $\mathrm{Law}[X \mid Z]$ gives the distribution of CAPE X within a latent state Z,
- $\mathrm{Law}[\#1, \#3 \mid Z]$ gives the joint probability distribution of number of grid points with positive and negative vertical velocity. For updraft, $\#1$ denotes $\#\{Y_i = 1\}$ and for downdraft, $\#3$ denotes $\#\{Y_i = 3\}$.

The missing number of neutral grid points $\#\{Y_i = 2\}$ follows from $\#2 = m - \#1 - \#3$ with m denoting the total number of grid points. **R2TC17:** In order to visualize the probabilities of the small scale conditioned on the latent states of the large scale variable, the entries of the λ matrix in (9) will be displayed dependent on the number of down- and updrafts. **R2SC38:** In Fig. 6, $K = 2$ bivariate histograms are shown for day and night respectively. Here the conditional probabilities of matrix λ are displayed for every latent state dependent on the number of up- and downdrafts (#1 and #3). Since the number of smaller-scale boxes is a_m, only the lower triangle below the diagonal corresponds to categories. Categories not populated by data are not shown (white). We noticed that in case the interval for vertical draft in Sect. 4.1 are increased, fewer data points are in the classifications for the up- and downdrafts (i.e. smaller numbers #1 and #3 change lower triangular probability matrices of Fig. 6). A comparison of different sizes of intervals for vertical draft is not shown here. Increasing the interval makes less up-/downdrafts, thus moving probability mass away from the diagonal, where large fractions of up-/downdrafts are sitting. In Fig. 6 the results are shown for a 4×4 grid, that means we have 16 grid boxes with vertical velocities. In the histograms the numbers of up- and downdraft range from 0 to 16. Variable Z_1 represents the latent state 'High', latent state Z_2 the state 'Low', as in Fig. 5. The latent states are stochastically disaggregated in probabilities which describe the chance of number of up- and downdrafts conditioned on the latent states 'High' and 'Low'. **R2SC40:** In the top left panel (Z_1, day) of Fig. 6 probability adds up for numbers of updrafts below 10 to 81 %. **R2SC41:** In the top right panel, much of the probability mass is allocated to states with no downdraft, and little updraft. For $\mathrm{Law}[\#1, \#3 \mid Z_2]$ in the bottom left panel for the day, high conditional

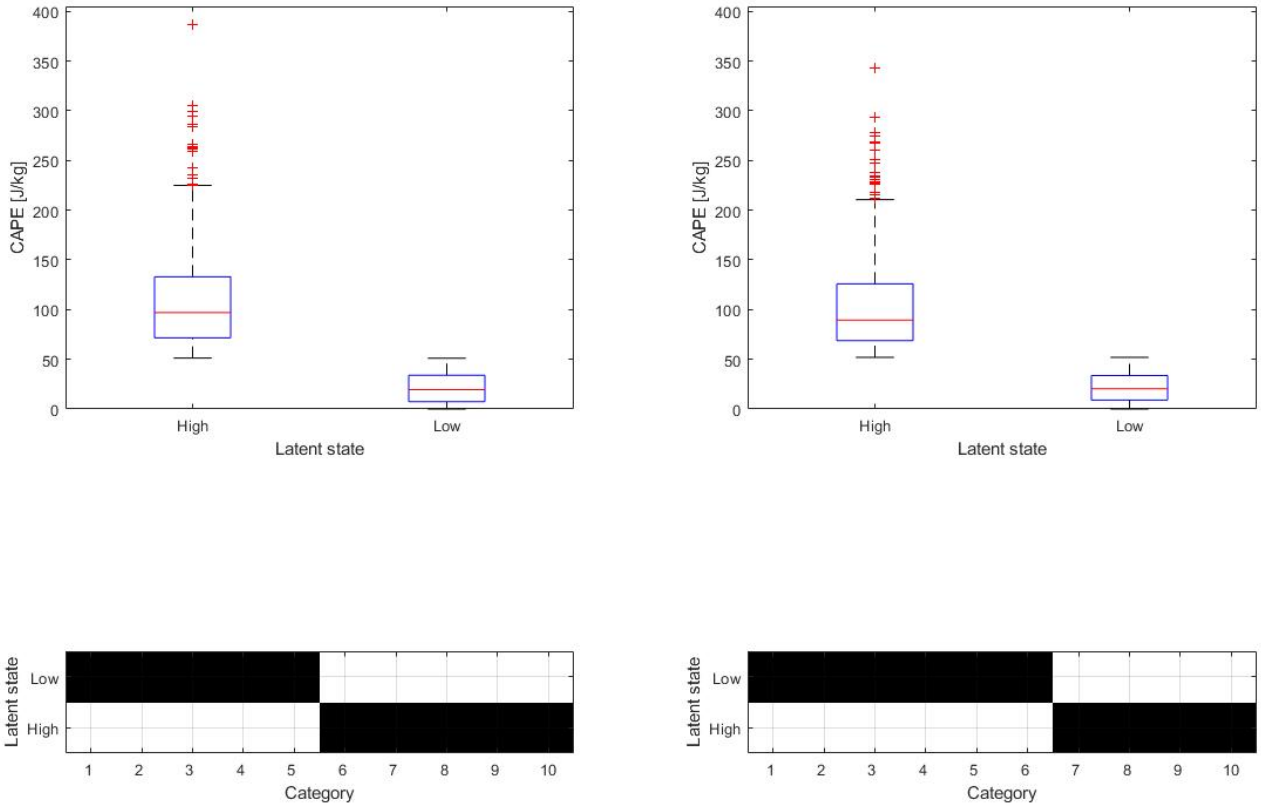

Figure 5. **R2SC36:** Top: Boxplots show the 12h averaged CAPE data which is affiliated to the latent states 'High' and 'Low'. The left boxplot presents 'Day' data and the right panel the 'Night' data; Bottom: Affiliation of CAPE categories to the latent states; CAPE data is spatially averaged over the Northwest quadrant of the COSMO-REA6 data and the vertical velocity is averaged for a box discretization of 64 grid boxes; see Tab. 1.

probabilities $\mathbb{P}[\#1, \#3 \mid Z_2]$ concentrate in categories with many boxes with downdraft. Here the probability of numbers of downdrafts of 6 to 16 is 68%. At night in the latent state Z_2, we observe that a low number of updraft boxes is likely, while the
280 overall up- and downdraft activity seems to be the least probable here (probability concentrating around $(\#1, \#3) \approx (0, 0)$).
R2SC42: In the bottom left panel (Z_2, night) the probability is accumulated to 82% for the number of updrafts between 0 and 4.

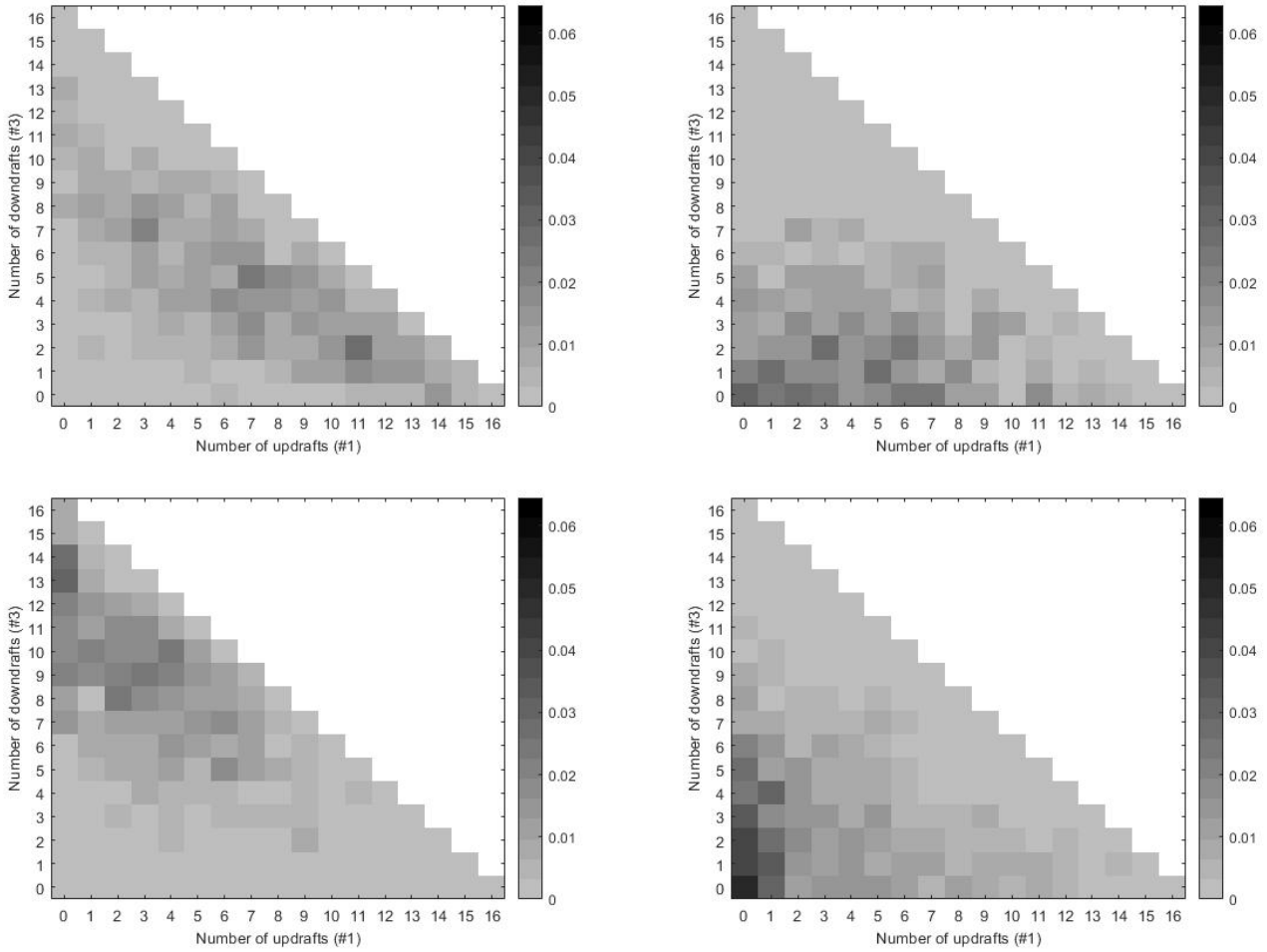

Figure 6. **R2SC39:** Histograms show probabilities of numbers of updrafts (#1) and downdrafts (#3) conditioned on latent states 'High' (top) and 'Low' (bottom); The left histograms present 'Day' data and the right histograms the 'Night' data; CAPE and vertical velocity data correspond to the data applied in Fig. 5.

4.1.3 Output on smaller scale

Note that the number of possible output categories $\hat{Y}$ scales quadratically with the number m of grid points considered on the smaller scale. Moving towards the convective scale, m increases, and so does the number of possible output categories, yet the number of data points (1302) stays the same. To avoid the resulting increase of estimation error, we further reduce the number of output categories by dividing the respective numbers for up- and downdraft into 3 sections, which leaves 6 categories. We use $15\,\mathrm{km} \times 15\,\mathrm{km}$ grid boxes on convective scale for the output of DBMR. The large scale remains unchanged compared to the previous example. In Fig. 7 the distribution of CAPE in terms of latent states based on kernel density estimation (KDE) is shown. At night, more categories are assigned to the latent state 'Low', the first latent state has a larger mean and median than during daytime.

In Fig. 8 the conditional probabilities are shown for 1024 boxes of vertical velocities. In the histograms the 3 sections of numbers of up- and downdraft range from 0 to 1024. The 3 categories are divided by the following numbers: 0 to 341, 342 to 683 and 684 to 1024 up- and downdrafts. Variable Z_1 represents again the latent state 'High', and Z_2 the latent state 'Low'; cf. Fig. 8. The first latent state is represented in the first row. During daytime down- or updraft is likely, and during nighttime it is most likely to have less downdraft than updraft. The smaller scale analysis gives consistent results with the analysis where the output is on mesoscale in Fig. 6. There are higher probabilities during daytime for medium to high numbers of up- or downdraft. At night due to less vertical draft, low to medium numbers of up- or downdraft are higher. For the second latent state 'Low', the distributions concentrate on higher and lower numbers of downdrafts and small numbers of updraft.

4.2 R2GC2: Higher number of latent states

The results for three latent states are considered in Appendix A. Figs. A1 and A2 show results using CAPE as input with a resolution of $500\,\mathrm{km} \times 500\,\mathrm{km}$ on large scale and a grid of $125\,\mathrm{km} \times 125\,\mathrm{km}$ for the output. The scale difference is again of factor 4 according to the first example in Sect. 4.1 where in- and output are on the synoptic scale. **R2SC45:** Affiliations without gaps lead to a separation of the latent states. "No gaps" means that there is no overlap and a clear separation of the latent states regarding the range of CAPE. This does not apply to every run with DBMR. Here we show the best ML bound estimate of 100 runs, see Fig. A1. The affiliations have no gaps for day and night. We have again more variance of the conditional probabilities during daytime. At night there is less variance of the conditional probabilities with a concentration at low numbers of downdraft or updraft boxes. A hierarchy of three different probability configurations arises for up-, down- and no draft. When the number of latent states K is further increased, the latent states can be clustered in groups of high, low and medium CAPE categories. In Fig. A1 top boxplots of CAPE categories by 3 latent states for daily mean (left: day and right: night) and in the middle the affiliation of CAPE categories to the latent states are presented. For higher K, the number of latent states with affiliation without gaps is higher at night compared to day.

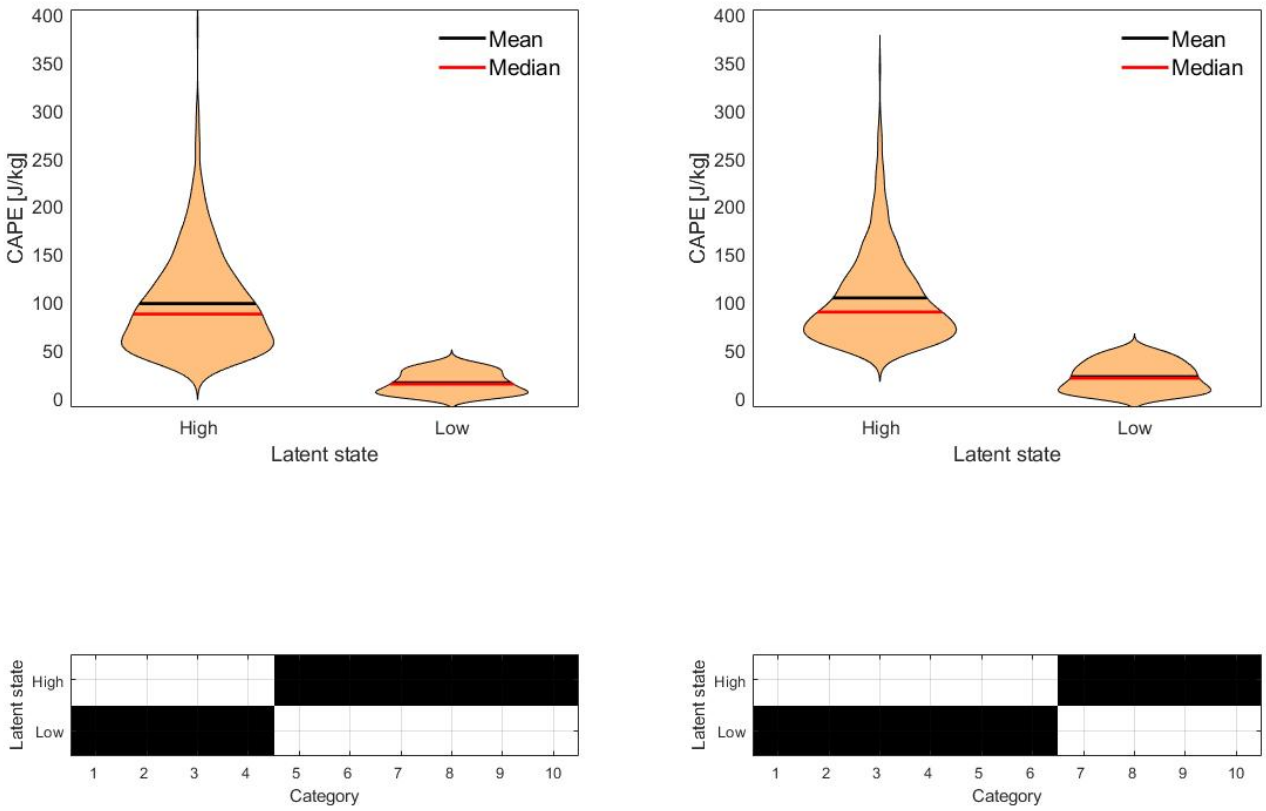

Figure 7. **R2GC2, R2SC43:** Top: Distributions based on kernel density estimation show the 12h averaged CAPE data which is affiliated to the latent states 'High' and 'Low'. The left boxplot presents 'Day' data and the right panel the 'Night' data; Bottom: Affiliation of CAPE categories to the latent states; CAPE data is spatially averaged over the Northwest quadrant of the COSMO-REA6 data and the vertical velocity is averaged for a box discretization of 4096 grid boxes; see Tab. 1.

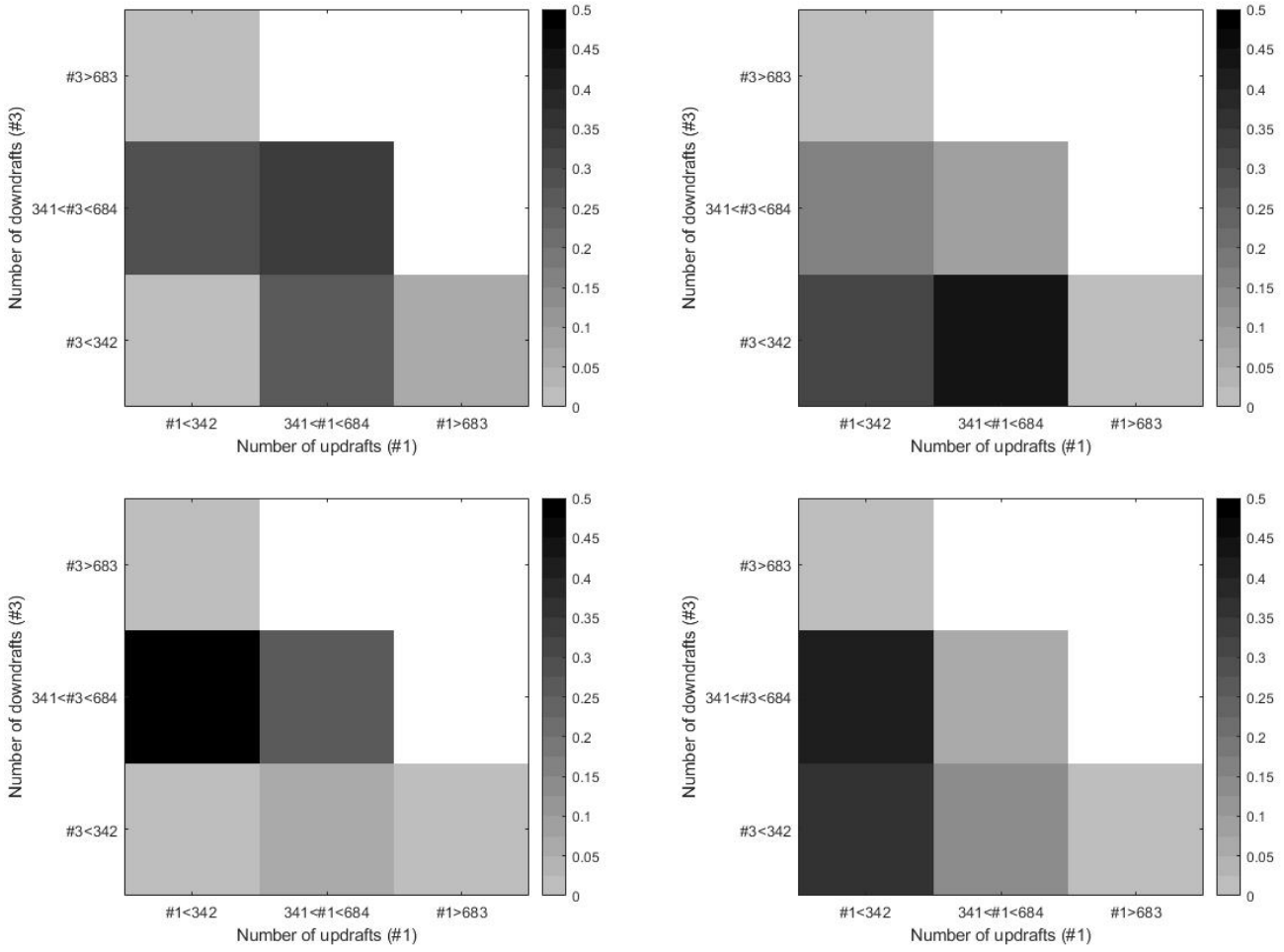

Figure 8. **R2GC2:** Histograms show probabilities of numbers of updrafts (#1) and downdrafts (#3) conditioned on latent states 'High' (top) and 'Low' (bottom); The left histograms present 'Day' data and the right histograms the 'Night' data; CAPE and vertical velocity data correspond to the data applied in Fig. 7.

4.3 R2GC2: Implications for general atmospheric dynamics

In Sect. 4.1 we discussed the results of the Bayesian model reduction from a mathematical perspective and in Sect. 4.2 we interpreted the outcomes for a higher number of latent states. The method groups input categories into fewer latent states. These are interpreted as reduced states for the large-scale atmospheric dynamics with respect to their probabilistic impact on vertical motion. We applied an energetic variable as the driver on large scale. CAPE is the convective available potential energy. It does not have to be fully available, meaning that high CAPE values does not *necessarily* lead to convective activity on smaller scales but increases the probability of smaller scale convective activity. The release of kinetic energy of a certain CAPE level to vertical movement needs triggers such as flows over mountains or forests which lead to instabilities of the hydrostatic equilibrium. The dependence on surface conditions on the earth requires a probabilistic way of thinking. Therefore the mathematical tool DBMR provides a simple probabilistic description. Using the method, we intend to draw conclusions about categorical processes in the atmosphere. Since the system can not be in two different categories simultaneously, categories are disjoint and the relation between the probability for large scale and smaller scales can be formulated via the conditional probabilities and the conservation of the total probability. The methodology breaks up probability calculations into distinct parts and relates marginal probabilities to conditional probabilities. The aim of this work is to test the stochastic method in an meteorological application towards a reduced categorical model of smaller scale convective activity in the atmosphere depending on large scale drivers.

To analyze the relation of large scale dynamics in the atmosphere to smaller scale categorical processes, the COSMO-REA6 reanalysis data set was applied; see Bollmeyer et al. (2015). We averaged CAPE for 500 km $\times$ 500 km and the vertical up- and downdrafts in 125 km $\times$ 125 km domains, as described in Sect. 3.2. Regarding the summer months July and August in the years 1995 to 2015, CAPE reaches averaged values between 0 and 400 J/kg and the vertical velocities have ranges from -0.15 to 0.2 m/s on mesoscale and -1.7 to 1 m/s on convective scale. In the meteorological setting we showed how the Bayesian model reduction performs. We combined large-scale CAPE with a subgrid-mesoscale time series for vertical velocity and count the numbers of up- and downdrafts. Therefore we mapped vertical velocities as updraft, no draft and downdraft dependent on an interval around zero vertical velocity. In the preprocessing of Sect. 4.1 we adjusted the interval for vertical draft with range 0.0096 m/s according to the meteorological data. The interval was chosen symmetrically on the basis of the histogram of mean vertical velocities in Fig. 3. We chose a number of 10 input categories and reduced these to two latent states. This was done for day and night, respectively.

In Fig. 5 the summary statistics with the affiliation of input categories to the latent states are presented. The affiliations in Figs. 5 and 7 have no gaps, meaning that the affiliations are interrelated and are not interrupted. The affiliations lead to a separation of the latent states in the boxplots for day and night. Thus a certain range of CAPE values can be assigned to every latent states. During daytime the range of values for the latent state ‘High’ is at around 400 J/kg and greater compared to the corresponding latent state during nighttime. For smaller scales we reduced the number of output categories. In Fig. 7 at the bottom, 6 high and 4 low CAPE categories for daily mean and 4 high and 6 low CAPE categories at night are affiliated. As a result of the averaging, the categories are almost evenly distributed over the latent states. The convective activity of the

atmosphere is stronger during the day than during nighttime. Therefore, the vertical draft is less at night than during the day. Mean and median are around 100 J/kg for the latent state 'High' and 25 J/kg for the latent state 'Low'. The mean and median are similar for day and night. There is a difference for the variance. At night the distribution of latent state 'High' is sharper due to less variance, only 4 categories are affiliated compared to the daily mean.

Joint probability distribution of number of grid points with positive and negative vertical velocity conditioned on the resulting latent states are shown in figs. 6 and 8. The sum of the probabilities of all categories for every box is 1. Increasing the interval for vertical draft makes less up-/downdrafts, thus moving probability mass away from the diagonal, where large fractions of up-/downdrafts are sitting. There are higher probabilities during daytime for medium to high numbers of up- or downdraft. Lots of updrafts during daytime lead to the existence of a lot of downdrafts due to mass conservation. At night due to less vertical draft, low to medium numbers of up- or downdraft are higher. For the latent state 'Low', the distributions in Figs. 6 and 8 concentrate on higher and lower numbers of downdrafts and small numbers of updraft. **R2SC46:** The representation of probabilities of numbers of updrafts and downdrafts conditioned on the latent states in Fig. 8 correspond in their distributions to the results on mesoscale in Fig. 6. The generation of kinetic energy of a certain CAPE level to vertical draft on smaller scales can occur up to a few hours later. A temporal shift for the in- and output could have an effect on the stochastic relation shown in Fig. 6. We consider the 12 hours means. For data with a higher temporal resolution, one could realize a shift of 2-4 hours for the input. This is deferred to future studies.

5 Conclusions

It is of importance to identify stochastic models by using categorical approaches compared to fluid mechanics described by continuous partial differential equations. In this study, a recent algorithmic framework called Direct Bayesian Model Reduction (DBMR) is applied which provides a scalable probability-preserving identification of reduced models directly from data; see Gerber and Horenko (2017). We assume that the output of a Bayesian model depends on the input through a latent variable, which can merely take a small number of different latent states. **R2SC47:** In this work, a direct Bayesian model reduction of smaller scale convective activity conditioned on large scale dynamics is investigated with regard to intermediate latent states. We combined the convective available potential energy (CAPE) as large scale flow variable with smaller scale subgrids time series for vertical velocity. Therefore we mapped vertical velocities as updraft, no draft and downdraft dependent on an interval around zero vertical velocity and count the numbers of up- and downdrafts. Data sets of daily means of 12 hours for day and night were computed using COSMO-REA6 reanalysis over a domain that covers Germany for a period of the summer months July and August in the years 1995 to 2015. In the analysis the scales from 500km to 125km (mesoscale) and up to 15km were considered. The categorical data analysis was done for day and night and discussed for different numbers of latent states. We chose a number of 10 input categories and reduced these to two and three latent states.

The step from the fluid continuum described by partial differential equations to a categorical stochastic description with DBMR provides a reduced model defined on a set of a few latent variables. These are interpreted as reduced states for the large scale atmospheric dynamics with respect to their probabilistic impact on vertical motion. For 2 latent states the input is

separated into categories with high and low CAPE values whereas for 3 latent states we have an affiliation to categories with high, medium and low CAPE values. The output categories for the vertical velocity describe the number of up- and downdrafts. In the result, we gain conditional distributions for the numbers of up- and downdrafts conditioned on the latent states for day and night. In the application we found a probabilistic relation of CAPE and vertical up- and downdraft.

For a resolution of 125km we applied a 4×4 grid and had 16 boxes with vertical velocities. During daytime the chance for updraft is higher conditioned on the latent state with high CAPE values. Probability adds up for numbers of up- or downdrafts higher than 10 to 81%. The distribution for the latent state with low CAPE values has higher probabilities at high numbers of downdrafts. Here the probability of numbers of downdrafts of 6 to 16 is 68%. At night probability adds up at small numbers of downdrafts for the latent state with high CAPE values. For low CAPE values, we observe that a low number of updrafts is likely. The probability is accumulated to 82% for the number of updrafts between 0 and 4.

On smaller scale with a resolution of 15km we applied a 32×32 grid and had 1024 boxes with vertical velocities. We divided the output into 3 categories of low (0 to 341), medium (342 to 683) and high (84 to 1024) numbers of up- and downdrafts. During daytime the probability for a medium number of up- and downdrafts is 34% for the latent state with high CAPE values. Here low and high numbers of up- and downdraft have small probability. For low CAPE values the maximum in the distribution occurs for a medium number of downdrafts and low number of updrafts at 50%. At night the probability adds up at low to medium numbers of downdrafts for the latent state with high CAPE values and for low CAPE values, we observe that the chance of low and medium number of updrafts is 82%. The distribution for the smaller scale resolution (15km) is a stochastic aggregation of the distribution with resolution of 125km. Therefore the distributions are qualitatively similar. When the number of latent states is further increased, the latent states can be clustered in groups of high, low and medium CAPE categories.

The model reduction of smaller scale convective activity is part of a development process for a model with a stochastic component for a conceptual description of convection embedded in a deterministic atmospheric flow model. Various energetic variable are applicable on large scale. A potential driver to control small scale models is the Dynamic State Index (DSI) in Müller et al. (2020) and Müller and Névir (2019), an "adiabaticity indicator". Other large scale variables driving the smaller scale stochastics are the available moisture or vertical wind shear. The presented approach provides a basis for further research of smaller scale convective activity conditioned on other possible large scale drivers.

Author contributions. Annette Müller prepared the meteorological data. All authors have then contributed to develop the work and prepared the manusscript.

Acknowledgements. This research has been funded by Deutsche Forschungsgemeinschaft (DFG) through grant CRC 1114 'Scaling Cascades in Complex Systems, Project Number 235221301, Project A01 'Coupling a multiscale stochastic precipitation model to large scale atmospheric flow dynamics'.

Appendix A

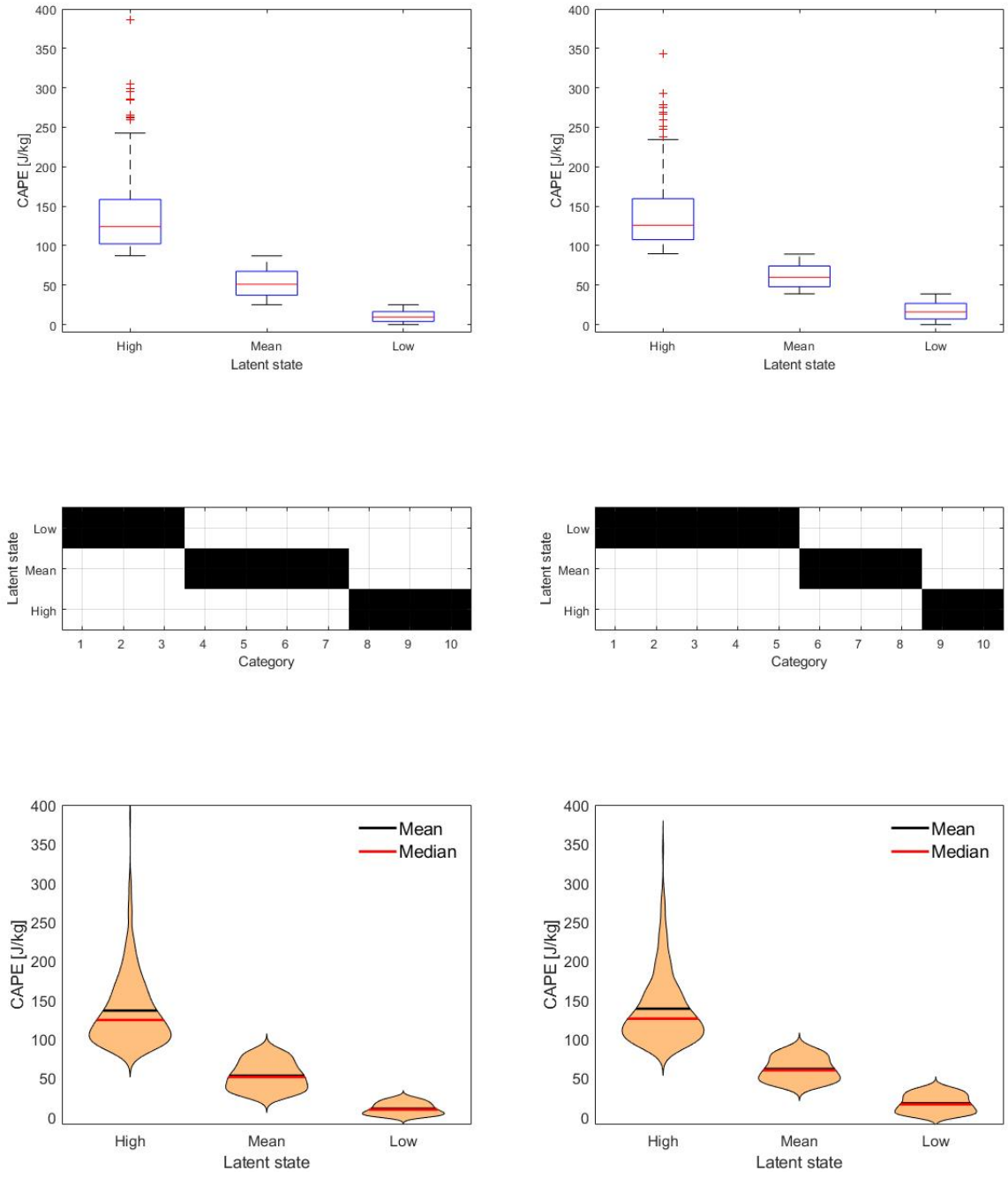

Figure A1. **R2GC2:** Boxplots show the 12h averaged CAPE data which is affiliated to the latent states 'High', 'Mean' and 'Low'. The left boxplot presents 'Day' data and the right panel the 'Night' data; Middle: Affiliation of CAPE categories to the latent states; Bottom: Distribution of CAPE in terms of latent states based on kernel density estimation; CAPE data is spatially averaged over the Northwest quadrant of the COSMO-REA6 data and the vertical velocity is averaged for a box discretization of 64 grid boxes; see Tab. 1.

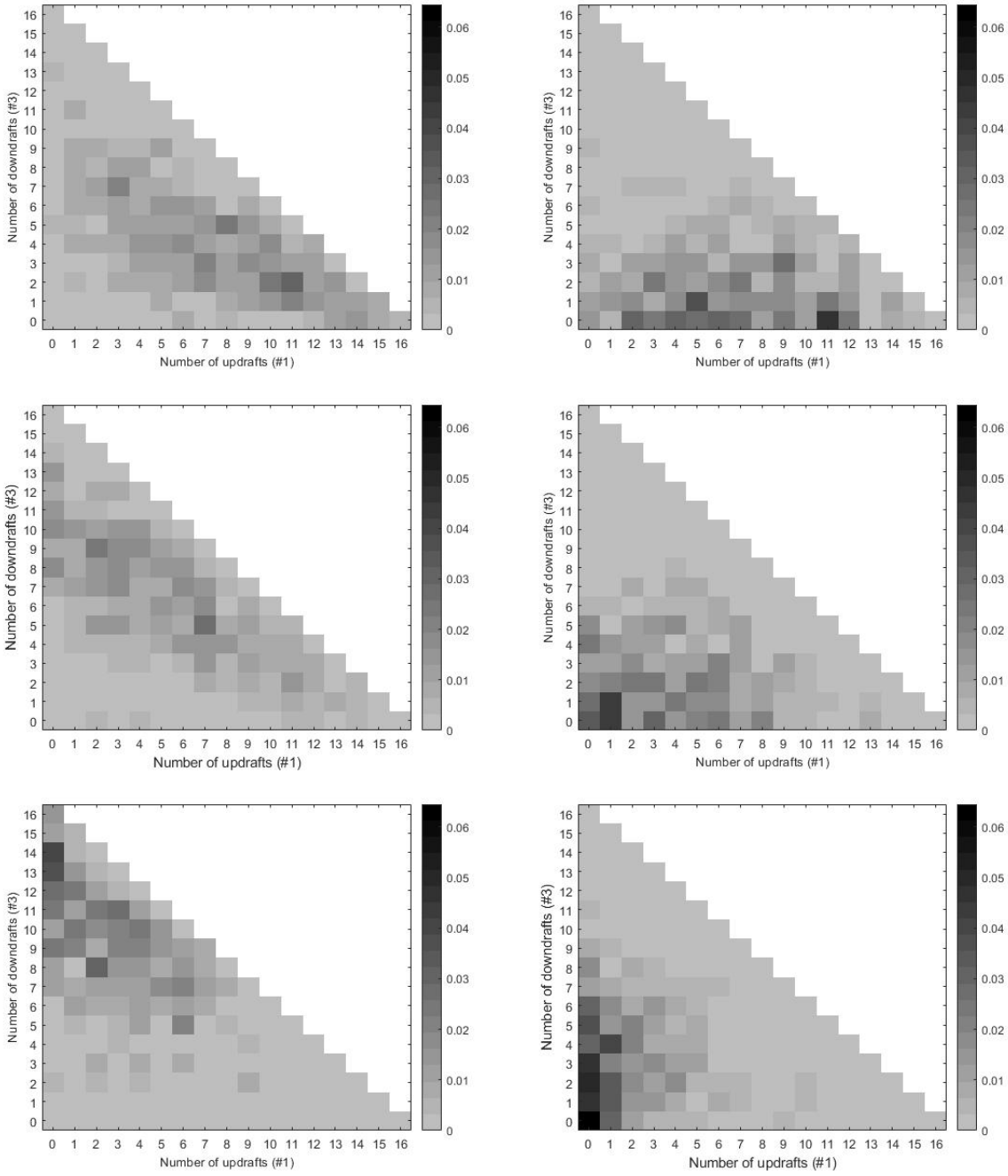

Figure A2. **R2GC2:** Histograms show probabilities of numbers of updrafts (#1) and downdrafts (#3) conditioned on latent states 'High' (top), 'Median' (middle) and 'Low' (bottom); The left histograms present 'Day' data and the right histograms the 'Night' data; CAPE and vertical velocity data correspond to the data applied in Fig. A1.

References

Berner, J., Achatz, U., Batte, L., Bengtsson, L., Cámara, A. d. l., Christensen, H. M., Colangeli, M., Coleman, D. R., Crommelin, D., Dolaptchiev, S. I., et al.: Stochastic parameterization: Toward a new view of weather and climate models, Bulletin of the American Meteorological Society, 98, 565–588, 2017.

Blanchard, D. O.: Assessing the vertical distribution of convective available potential energy, Weather and Forecasting, 13, 870–877, 1998.

Bollmeyer, C., Keller, J., Ohlwein, C., Wahl, S., Crewell, S., Friederichs, P., Hense, A., Keune, J., Kneifel, S., Pscheidt, I., et al.: Towards a high-resolution regional reanalysis for the European CORDEX domain, Quarterly Journal of the Royal Meteorological Society, 141, 1–15, 2015.

Bott, A.: Synoptische Meteorologie: Methoden der Wetteranalyse und-prognose, Springer-Verlag, 2016.

Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F., and Zucker, S. W.: Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps, Proceedings of the national academy of sciences, 102, 7426–7431, 2005.

Donoho, D. L. and Grimes, C.: Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data, Proceedings of the National Academy of Sciences, 100, 5591–5596, 2003.

Dorrestijn, J., Crommelin, D., Biello, J., and Böing, S.: A data-driven multi-cloud model for stochastic parametrization of deep convection, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 371, 20120 374, 2013a.

Dorrestijn, J., Crommelin, D. T., Siebesma, A. P., and Jonker, H. J.: Stochastic parameterization of shallow cumulus convection estimated from high-resolution model data, Theoretical and Computational Fluid Dynamics, 27, 133–148, 2013b.

Dutton, J.: Dynamics of Atmospheric Motion, 617 pp, 1976.

Franzke, C. L., O'Kane, T. J., Berner, J., Williams, P. D., and Lucarini, V.: Stochastic climate theory and modeling, Wiley Interdisciplinary Reviews: Climate Change, 6, 63–78, 2015.

Fritsch, J. and Chappell, C.: Numerical prediction of convectively driven mesoscale pressure systems. Part I: Convective parameterization, Journal of Atmospheric Sciences, 37, 1722–1733, 1980.

Gerber, S. and Horenko, I.: Toward a direct and scalable identification of reduced models for categorical processes, Proceedings of the National Academy of Sciences, 114, 4863–4868, 2017.

Gerber, S., Olsson, S., Noé, F., and Horenko, I.: A scalable approach to the computation of invariant measures for high-dimensional Markovian systems, Scientific reports, 8, 1796, 2018.

Gottwald, G. A., Peters, K., and Davies, L.: A data-driven method for the stochastic parametrisation of subgrid-scale tropical convective area fraction, Quarterly Journal of the Royal Meteorological Society, 142, 349–359, 2016.

Holland, P. W.: Statistics and causal inference, Journal of the American statistical Association, 81, 945–960, 1986.

Horenko, I.: On simultaneous data-based dimension reduction and hidden phase identification, Journal of the atmospheric sciences, 65, 1941–1954, 2008.

Horenko, I., Dolaptchiev, S. I., Eliseev, A. V., Mokhov, I. I., and Klein, R.: Metastable decomposition of high-dimensional meteorological data with gaps, Journal of the atmospheric sciences, 65, 3479–3496, 2008.

Jolliffe, I.: Principal component analysis, Technometrics, 45, 276, 2003.

Khouider, B., Biello, J., Majda, A. J., et al.: A stochastic multicloud model for tropical convection, Communications in Mathematical Sciences, 8, 187–216, 2010.

Kirkpatrick, C., McCaul Jr, E. W., and Cohen, C.: Variability of updraft and downdraft characteristics in a large parameter space study of convective storms, Monthly Weather Review, 137, 1550–1561, 2009.

Klein, R.: Scale-dependent models for atmospheric flows, Annual review of fluid mechanics, 42, 249–274, 2010.

Lorenz, E. N.: Empirical orthogonal functions and statistical weather prediction, 1956.

Moncrieff, M. W. and Miller, M. J.: The dynamics and simulation of tropical cumulonimbus and squall lines, Quarterly Journal of the Royal Meteorological Society, 102, 373–394, 1976.

Müller, A. and Névir, P.: Using the concept of the Dynamic State Index for a scale-dependent analysis of atmospheric blocking, Meteorologische Zeitschrift, 28, 487–498, https://doi.org/10.1127/metz/2019/0963, 2019.

Müller, A., Niedrich, B., and Névir, P.: Three-dimensional potential vorticity structures for extreme precipitation events on the convective scale, Tellus A: Dynamic Meteorology and Oceanography, 72, 1–20, 2020.

Schmid, P. J.: Dynamic mode decomposition of numerical and experimental data, Journal of fluid mechanics, 656, 5–28, 2010.

Schölkopf, B., Smola, A., and Müller, K.-R.: Kernel principal component analysis, in: International conference on artificial neural networks, pp. 583–588, Springer, 1997.

Von Luxburg, U.: A tutorial on spectral clustering, Statistics and computing, 17, 395–416, 2007.

Weisman, M. L. and Klemp, J. B.: The dependence of numerically simulated convective storms on vertical wind shear and buoyancy, Monthly Weather Review, 110, 504–520, 1982.

Zhao, Y., Levina, E., Zhu, J., et al.: Consistency of community detection in networks under degree-corrected stochastic block models, The Annals of Statistics, 40, 2266–2292, 2012.