Modeling long correlation times using additive binary Markov chains: Applications to wind generation time series

Juliane Weber,[1,2] Christopher Zachow,[2] and Dirk Witthaut[1,2,*]

[1]*Institute of Energy and Climate Research–Systems Analysis and Technology Evaluation, Forschungszentrum Jülich, 52425 Jülich, Germany*
[2]*Institute for Theoretical Physics, University of Cologne, Zülpicher Straße 77, 50937 Cologne, Germany*

Wind power generation exhibits a strong temporal variability, which is crucial for system integration in highly renewable power systems. Different methods exist to simulate wind power generation but they often cannot represent the crucial temporal fluctuations properly. We apply the concept of additive binary Markov chains to model a wind generation time series consisting of two states: periods of high and low wind generation. The only input parameter for this model is the empirical autocorrelation function. The two-state model is readily extended to stochastically reproduce the actual generation per period. To evaluate the additive binary Markov chain method, we introduce a coarse model of the electric power system to derive backup and storage needs. We find that the temporal correlations of wind power generation, the backup need as a function of the storage capacity, and the resting time distribution of high and low wind events for different shares of wind generation can be reconstructed.

I. INTRODUCTION

Mitigation of climate change requires the decarbonization of the energy system [1–3]. Power plants based on fossil fuels must be replaced by renewable sources such as wind and solar power. These technologies have shown remarkable progress in recent decades, but the integration into the energy system represents a huge challenge due to their strong intermittency [4–10]. Generation and load have to be balanced at every instance of time. Thus, for a system dominated by fluctuating renewable sources, large amounts of backup and storage are required in order to always guarantee this power balance [11–14].

Wind power, as one of the main renewable energy sources, exhibits temporal correlations on different time scales from seconds [15,16] over weeks and years [17] to decades [5]. Periods with below (and above) average wind power generation can last up to weeks [18,19]. Furthermore, the generation can vary significantly on hourly time scales, even when aggregated over large spatial scales [20–23]. These fluctuations have to be accounted for when designing a future energy system.

High amounts of backup and storage infrastructures will be necessary in energy systems with a high share of wind power to provide energy in times of low renewable generation [11,12,24,25]. In order to quantify the need for backup and storage, either long-term measurements or reliable models of wind power generation are needed. As long-term measurements are rare, it is important to develop models which are able to represent the temporal fluctuations of wind power generation properly.

Stochastic models are commonly used to simulate wind speed and wind generation data [20,26–30]. Typical models include autoregressive moving average and Markov chain models. Autoregressive moving average models require a high amount of model parameters. Simple first-order Markov models often cannot reproduce the temporal characteristics of wind power generation, in particular long correlation times, and therefore they are of very limited use for energy system analyses [26]. Higher-order Markov models require a huge amount of input data and are thus impractical [20,26,27]. Additionally, for ordinary Markov models the data have to be discretized and the model output depends on the exact discretization scheme [20].

Stochastic models for wind power generation should reproduce the main temporal characteristics but remain simple enough for practical applications. We introduce a simple two-state stochastic model based on additive binary Markov chains developed in [31–34]. Additive Markov chains are an efficient tool to simulate time series with long-range correlations because the transition probability can be expressed as a sum of functions, each depending on one of the previous states (memory functions). Strong analytic results exist for the case of binary time series [31–34]. In this case, the memory functions can be derived from the empirical autocorrelation function of the time series straightforwardly. Additionally, the model depends on the memory length which determines the maximum time lag taken into account to derive the memory functions. In this paper we thus consider a simplified, binary wind generation time series, distinguishing only between two different types of system states: scarcity in times of low and oversupply in times of high wind generation. Already for this simple model, in which the wind power generation can assume only two values for the two states, we obtain a fairly good stochastic model. Our purpose is to reconstruct the temporal aspects of a wind power time series as they are crucial for system operation, especially for high shares of renewables combined with large storage facilities. Therefore, we evaluate our model by analyzing the temporal correlations, the backup need as a function of the storage capacity, and the resting time

*d.witthaut@fz-juelich.de

distributions of scarcity and overproduction events for different shares of wind generation.

The paper is organized as follows. We introduce the additive binary Markov chain method (Sec. II), discuss the available wind power data, and present a coarse-grained model of the electric power system used to test the stochastic model (Sec. III). Subsequently, we compare the characteristics of the input and the synthetic data in order to evaluate our model (Sec. IV). In a further step, we present a way to extend our model to nonbinary wind generation data (Sec. V). We close with some concluding remarks (Sec. VI).

II. ADDITIVE BINARY MARKOV CHAINS

A one-step Markov chain of a homogeneous stochastic process with a discrete number of states is defined as a sequence of random variables $X_1, X_2, X_3, \ldots$ in which the conditional probability for the future state $X_t = x_t$ is determined entirely by the knowledge of the previous state $X_{t-1} = x_{t-1}$ [35,36],

$$
\begin{aligned}
P(X_t = x_t | X_{t-1} &= x_{t-1}, \ldots, X_0 = x_0) \\
&= P(X_t = x_t | X_{t-1} = x_{t-1}).
\end{aligned}
\tag{1}
$$

For practical applications, this conditional probability is reconstructed from measured data by discretizing the data such that the state space of the random variable becomes finite.

To include memory, Eq. (1) can be generalized to an N-step Markov chain for which the future state depends on N previous states,

$$
\begin{aligned}
P(X_t = x_t | X_{t-1} &= x_{t-1}, \ldots, X_0 = x_0) \\
&= P(X_t = x_t | X_{t-1} = x_{t-1}, \ldots, X_{t-N} = x_{t-N}).
\end{aligned}
\tag{2}
$$

However, as Brokish and Kirtley [26] point out, "the problem with higher-order Markov models is that there are K^N states, where K is the number of discretized wind powers and N is the order of the model, which is intractable for large N" and therefore for long memory times.

Assuming that the influence of previous states on the future state is additive, Eq. (2) can be expressed as a sum of functions f, each depending on one of the previous states

$$
\begin{aligned}
P(X_t = x_t | X_{t-1} &= x_{t-1}, \ldots, X_{t-N} = x_{t-N}) \\
&= \sum_{r=1}^{N} f(x_t, x_{t-r}, r).
\end{aligned}
\tag{3}
$$

Strong analytical results are available for additive binary Markov chains, i.e., stochastic processes which assume only two states $a_t = \{0, 1\}$ [31–33]. In this case, the conditional probability can be simplified to [31–33]

$$
P(a_t = 1 | T_{N,t}) = \sum_{r=1}^{N} f(a_{t-r}, r),
\tag{4}
$$

with $T_{N,t} = a_{t-1}, a_{t-2}, \ldots, a_{t-N}$. The function $f(a_{t-r}, r)$ describes the contribution of a_{t-r} to the conditional probability of a_t taking the value $a_t = 1$ at time step t. The value N can be interpreted as the memory length of the Markov chain. Melnyk

et al. [33] show that Eq. (4) can be rewritten in the simpler form

$$
P(a_t = 1 | T_{N,t}) = \langle a \rangle + \sum_{r=1}^{N} F(r)(a_{t-r} - \langle a \rangle),
\tag{5}
$$

with $\langle a \rangle$ the average over the whole sequence and $F(r)$ a memory function, which describes the strength of the influence of the previous value a_{t-r} on a_t.

A relationship between the memory function $F(r')$ and the autocovariance function $K(r)$ exists [33],

$$
K(r) = \sum_{r'=1}^{N} F(r') K(r - r'), \quad r \geqslant 1,
\tag{6}
$$

with

$$
K(r) = \langle a_t a_{t+r} \rangle - \langle a \rangle^2,
\tag{7}
$$

$$
K(0) = \langle a \rangle (1 - \langle a \rangle),
\tag{8}
$$

$$
K(r) = K(-r).
\tag{9}
$$

Equation (6) is a system of linear equations and can be solved for $F(r')$ straightforwardly. Thus, it is sufficient to know the autocovariance function of a time series in order to reconstruct the memory function which fully characterizes the stochastic process. This memory function can then be inserted into Eq. (5) to compute synthetic time series for Monte Carlo simulations.

III. APPLICATION TO WIND GENERATION DATA

We apply the additive binary Markov model introduced in Sec. II to wind generation time series. One prime benchmark for such a model is whether important characteristics of power system operation can be reproduced. For this purpose, we introduce a simple coarse-grained model of the electric power system in Sec. III A to derive backup and storage needs which crucially depend on the temporal characteristics of wind power generation, especially for high shares of wind power [11,12]. Afterwards, we present the empirical data used as input for the construction of the model (Sec. III B) and explain the mapping of the continuous wind generation time series to a binary time series (Sec. III C). Finally, we explain the initialization of our model in Sec. III D.

A. Model of the electric power system

At every time step t, the generated and consumed electric power must be balanced [11,12,37]:

$$
R(t) + B(t) = \Delta(t) + L(t) + C(t).
\tag{10}
$$

Here $R(t)$ is the renewable power generation time series which we want to model using the additive binary Markov chain method and $L(t)$ is the load. If the load exceeds the generation, i.e., $L(t) > R(t)$, the missing energy must be provided by storage facilities [$\Delta(t) < 0$] or by conventional backup power plants [$B(t) > 0$]. The backup need can be interpreted as the aggregated amount of power required from dispatchable power plants. If, on the other hand, the generation exceeds the load, i.e., $R(t) > L(t)$, either excess energy can be stored [$\Delta(t) > 0$] or the generation from renewables has to be curtailed

$[C(t) > 0]$. Hence, the storage filling level $S(t)$ evolves according to

$$S(t + T) = S(t) + \Delta(t)T, \tag{11}$$

where T denotes the duration of one time step (here 1 h). Storage facilities are limited such that the storage filling level must satisfy $0 \leqslant S(t) \leqslant S_{\max}$.

The energy that has to be provided by conventional backup power plants is minimized for a storage-first strategy [12]. Thus, $\Delta(t)$ can be derived in the following way:

$$\Delta(t) = \begin{cases} \min\left[R(t) - L(t); \frac{S_{\max} - S(t)}{T}\right] & \text{if } R(t) > L(t) \\ -\min\left[L(t) - R(t); \frac{S(t)}{T}\right] & \text{if } R(t) < L(t). \end{cases} \tag{12}$$

Accordingly, $B(t)$ and $C(t)$ read

$$B(t) = \begin{cases} L(t) - R(t) + \Delta(t) & \text{if } L(t) > R(t) \\ 0 & \text{otherwise,} \end{cases} \tag{13}$$

$$C(t) = \begin{cases} R(t) - L(t) - \Delta(t) & \text{if } R(t) > L(t) \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

A central question for energy system operation is how much storage and backup are needed in a highly renewable power system [11,12,24]. We therefore use the temporal mean of the backup energy $\langle B \rangle$ as a function of the storage capacity $S_{\max}$ as a benchmark for the developed Markov chain model. For simplicity, we normalize the average backup energy by the average load $\langle L \rangle$ to obtain the average share of energy that has to be provided by backup power plants:

$$E = \frac{\langle B \rangle}{\langle L \rangle}. \tag{15}$$

A feature of the operation of the storage system is whether we are in the regime of scarcity $[R(t) < L(t)]$ or overproduction $[R(t) > L(t)]$. Hence, it is reasonable to consider the wind generation time series as a two-state stochastic system (cf. Fig. 1, described further in Sec. III C). In this case, we can make use of the relation between the autocovariance function $K(r)$ of the time series and the memory function $F(r)$ [Eq. (6)] such that it is straightforward to deduce the stochastic model from measured data.

B. Data

We apply the additive binary Markov model to generate synthetic wind generation time series. As input we use the renewables.ninja [38] wind data set for Germany, which consists of hourly wind capacity factors (i.e., wind power normalized by the rated capacity) based on MERRA reanalysis data [39] simulating the 2014 fleet of wind farms for 1985–2014. A wind generation time series exhibits a seasonal as well as a diurnal periodicity [20,29,30,40,41]: The generation in Germany is usually higher in winter than in summer. The diurnal wind power variation weakly depends on the respective season. Furthermore, the wind generation time series exhibits "good" and "bad" wind years as well as fluctuations on the synoptic time scale (which is of the order of a few days). Therefore, a wind generation time series is not a stationary process. However, Thomann and Barfield [42] argue that the effect of nonstationarity can be neglected as long as the time series does not exhibit a trend. The latter statement is true because a fixed installed capacity of wind farms was used in order to generate the renewables.ninja [38] data set. Additionally, it is valid to assume that the average wind generation does not exhibit a trend as long as long time frames are considered. As we use 30 years of hourly wind data, this condition is fulfilled.

As our model cannot represent the deterministic long-range seasonal variations, we seasonally adjust the input data by multiplicatively removing the typical seasonal variation of the wind generation time series using monthly averages. For this purpose, we divide the data of a respective month $R_{\mathrm{month}}(t)$ by a corresponding seasonality factor sf_{month} given as

$$sf_{\mathrm{month}} = \frac{\langle R_{\mathrm{month}}(t) \rangle}{\langle R(t) \rangle}, \tag{16}$$

where $\langle \cdot \rangle$ denotes the average over either all data corresponding to one month or the whole time frame.

We consider different scenarios for the expansion of wind power in the system. Therefore, we scale the capacity factor time series $w(t)$ from renewables.ninja [38] such that the average generation by wind power plants provides a certain share γ of the average load [11,12], i.e.,

$$R(t) = \gamma \frac{w(t)}{\langle w \rangle} \langle L \rangle. \tag{17}$$
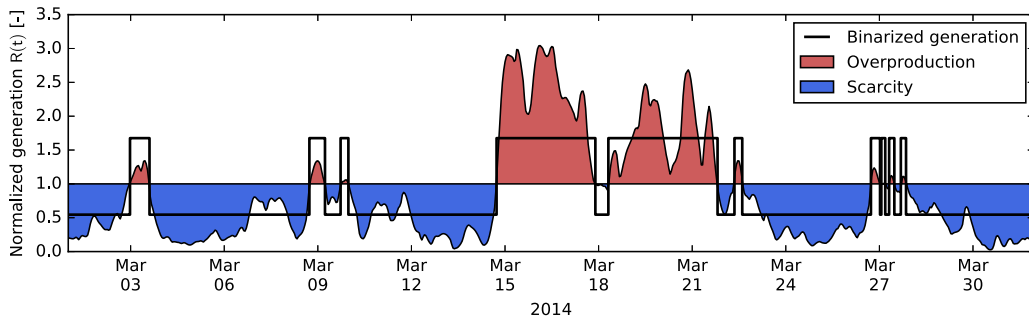

FIG. 1. Normalized, deseasonalized wind generation $R(t)$ (thin black line) and binarized time series (thick black line). The values of the two states R_0 and R_1 are derived from the full renewables.ninja [38] time series [cf. Eqs. (19) and (20)]. The load time series is normalized to $L(t) \equiv \langle L \rangle \equiv 1$ (see the text) and the wind generation time series is scaled with $\gamma = 1$ such that $\langle R \rangle = 1$ [cf. Eq. (17)]. Therefore, red shaded areas represent periods of overproduction $[R(t) \geqslant 1]$ and blue shaded areas represent periods of scarcity $[R(t) < 1]$.

The factor γ also denotes renewable penetration. In the case of a fully renewable power system, which we want to consider here primarily, γ is set to one. Other scenarios are also discussed (Sec. IV C).

For simplicity, we assume a constant load time series normalized to one [$L(t) \equiv \langle L \rangle \equiv 1$] in the following. In the case of a highly renewable power system, which we consider in this paper, the fluctuations of the wind generation time series are much higher than those of the load time series. Therefore, setting L to a constant value has only a minor influence on backup and storage needs. Furthermore, our primary objective is to evaluate the performance of the additive binary Markov chain model which is facilitated by simplifying the load time series.

C. Binarization of the data

The task of this paper is to develop a simple stochastic model for the intermittent wind time series $R(t)$, hence we want to represent the stochastic hour-by-hour fluctuations, the diurnal cycle, and variations on the synoptic time scale. To harness the strength of additive binary Markov chains, we need to define a reasonable mapping from the continuous values of $R(t)$ to a binary time series $a(t)$ with values zero and one. One example for such a mapping is to define a threshold value which separates the wind power generation into regimes of underproduction [$R(t) < \langle L \rangle$] and overproduction [$R(t) \geqslant \langle L \rangle$]

$$a(t) = \begin{cases} 0 & \text{if } R(t) < \langle L \rangle \\ 1 & \text{if } R(t) \geqslant \langle L \rangle. \end{cases} \tag{18}$$

From the binarized time series $a(t)$ we can derive the autocovariance function $K(r)$ [cf. Eq. (7)]. Choosing an appropriate memory length N, the memory function $F(r)$ can be recovered using Eq. (6) and the Markov chain can be simulated using Eq. (5).

In order to apply the power balance equation (10) to the binarized time series, we map $a(t)$ to average values of underproduction and overproduction of the input time series (see Fig. 1),

$$a(t) = 0 \Rightarrow R_{\text{bin}}(t) = R_0 = \langle R \rangle|_{R(t) < \langle L \rangle}, \tag{19}$$

$$a(t) = 1 \Rightarrow R_{\text{bin}}(t) = R_1 = \langle R \rangle|_{R(t) \geqslant \langle L \rangle}, \tag{20}$$

where $\langle \cdot \rangle|_x$ is the conditional expectation value. As we want to fix the renewable penetration γ for all simulations, the simulated values for R_0 and R_1 have to be scaled further such that $\langle R \rangle = \gamma$ always holds [cf. Eq. (17)].

D. Initialization of the model

The synthetic time series are modeled in such a way that the original and the simulated time series are of the same length (30 years in hourly resolution). In each case, an ensemble of 100 Monte Carlo simulations is considered.

The first N elements in the synthetic sequence are generated using the following modification of Eq. (5):

$$P(a_1 = 1) = \langle a \rangle$$
$$\text{if } t = 1, \tag{21}$$

$$P(a_t = 1 | T_t) = \langle a \rangle + \sum_{r=1}^{t} F(r)(a_{t-r} - \langle a \rangle)$$
$$\text{if } 1 < t \leqslant N. \tag{22}$$

This is considered to be a reasonable assumption, as the first N values represent only a very small fraction of the entire simulated stochastic sequence.

IV. RESULTS

In the following sections we discuss the memory function of the binarized time series used in this paper (Sec. IV A) and subsequently evaluate our additive binary Markov chain model by analyzing the autocorrelation function, the backup and storage need, and the resting time distribution for different memory lengths $N \in \{7, 14, 21, 70\} \times 24$ h (in the following, h means hours and d means days) in comparison to a simple memoryless Markov chain (Sec. IV B). Additionally, we assess the impact of the threshold value used to binarize the input data (Sec. IV C) by choosing different values for the renewable penetration γ.

A. Memory function

In order to simulate binary wind generation data using the additive binary Markov model, a memory function $F(r)$ has to be derived by solving the set of linear equations (6) for a predefined memory length N. In Fig. 2 the memory function $F(r)$ is shown for memory lengths $N \in \{7, 14, 21, 70\} \times 24$ h. The memory function is hardly sensitive to small changes of N. For all four choices of N, it takes a value of 0.90 for $r = 1$ h and is close to zero for $r > 1$ h (shown in detail in the inset in Fig. 2). The inset further shows that $F(r)$ decreases monotonically up
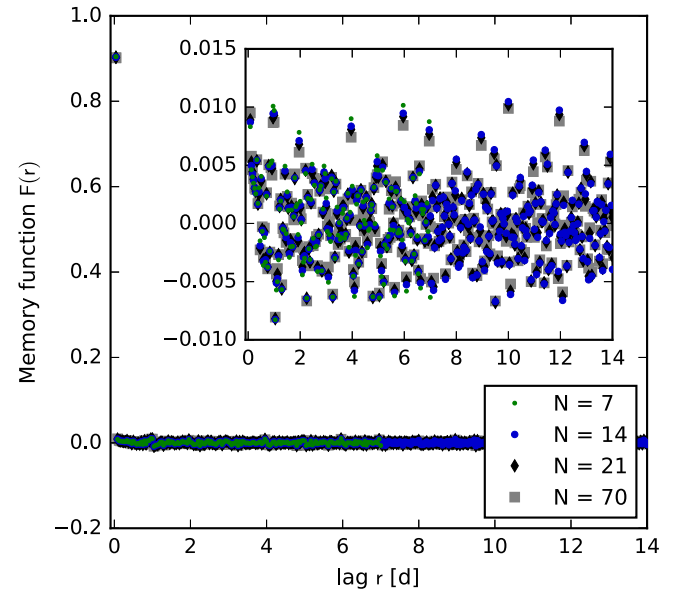
FIG. 2. Memory function $F(r)$ for memory lengths of $N \in \{7, 14, 21, 70\}$ d (green dots, blue circles, black diamonds, and gray boxes, respectively). The inset shows a zoom into the region around $F(r) = 0$ for better visualization. The time lag r is given in hourly resolution.
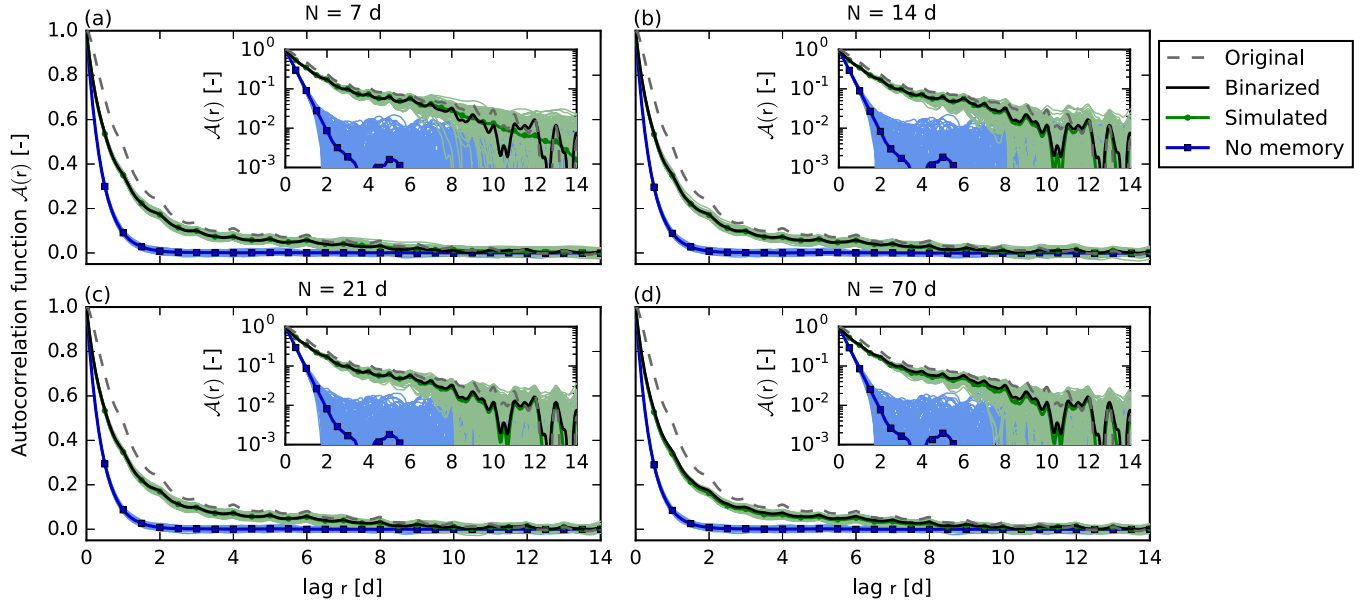
FIG. 3. Autocorrelation functions of the original renewables.ninja [38] (gray dashed lines) and the binarized (black solid lines) time series and of an ensemble of 100 simulations using the additive binary Markov chain approach for memory lengths of $N \in \{7,14,21,70\}$ d (green dots) and without memory, i.e., $F(r > 1) = 0$ (blue squares). The ensemble averages are represented as thick green and blue lines, respectively. For better visualization of the long-time decay, the autocorrelation functions are also shown on a logarithmic y scale in the insets.

to $r = 7$ h (and nonmonotonically up to $r \approx 12$ h) and exhibits peaks corresponding to a diurnal circle, according to the diurnal variation of the input time series.

The high difference in the magnitude of $F(1)$ and $F(r > 1)$ imposes the question if it is possible to obtain the same results when only using $F(1)$ for the simulation. This would imply that the stochastic process is memoryless and can be described by a simple one-step binary Markov chain. To evaluate this, we also simulate data using only $F(1)$ in Eq. (5), setting $F(r > 1) = 0$. In the following, we thus compare the results of the additive binary Markov model with memory to (i) the empirical binary time series and (ii) the results of the binary Markov chain without memory.

B. Additive Markov chain of the binarized wind power data

1. Autocorrelation function

The autocorrelation function $\mathcal{A}(r)$, i.e., the autocovariance function $K(r)$ normalized such that $\mathcal{A}(0) = 1$, describes how strong the wind generation at one time step is correlated to the wind generation at a different, lagged time step ($r > 0$). We observe that for the original as well as for the binarized time series, the autocorrelation function decreases fast for time lags up to about $r = 3$ d and levels off for higher time lags, being close to zero for about $r > 10$ d (gray dashed and black solid lines in Fig. 3). Additionally, the curves are superimposed by a slight 24-h oscillation, representing the diurnal variation of the time series. The binarization of the original time series leads to an underestimation of the autocorrelation function for all lags r. This can be explained in the following way. In the binarized time series, the wind generation can only take two values, zero and one. Therefore, the absolute value of the correlation [cf. Eq. (7)] is either high, i.e., $a(t + r) = a(t)$, or low, i.e., $a(t + r) \neq a(t)$. Small changes are neglected. Because of this, the

autocorrelation function of the binarized time series $\mathcal{A}_{\text{bin}}$ must be used as a reference in order to evaluate the additive binary Markov chain method.

The ensemble average of the autocorrelation functions derived from the additive binary Markov chain with memory lengths $N \in \{7,14,21,70\}$ d, $\mathcal{A}_{\text{Markov}}$ (thick green line in Fig. 3), is almost identical to the empirical $\mathcal{A}_{\text{bin}}$ for all lags r and all memory lengths N. Even the diurnal oscillation can be recovered. The autocorrelation function can only be reconstructed for time lags up to N as the conditional probability $P(a_t = 1|T_{N,t})$ in Eq. (5) directly depends on N. For higher time lags, the fluctuations in the autocorrelation function cannot be reproduced any longer [cf. Fig. 3(a)]. Furthermore, for high memory lengths, i.e., $N > 21$ d, the autocorrelation function tends to be slightly underestimated for all values of the time lag r. Hence, the memory length N must be chosen long enough to capture all significant correlations in the original binarized data, but not longer. Indeed, choosing a medium value of about $N = 7, \ldots, 21$ d, the additive binary Markov model reproduces the correlation structure of the binarized wind generation time series almost perfectly.

Without memory, i.e., $F(r > 1) = 0$ (blue lines in Fig. 3), the autocorrelation function is strongly underestimated. The correlation approaches zero for time lags of $r > 2$ d. Hence, without memory, the long-range correlations of the wind generation time series cannot be represented. We conclude that even though the values of $F(r)$ are small for $r > 1$ h (cf. Fig. 2), they still contain important information for the simulation.

2. Backup and storage need

A reliable wind power model should reproduce the essential characteristics of the original (binarized) time series. A key benchmark for power system operation is the average amount of backup energy E as a function of the storage capacity S_{max}.
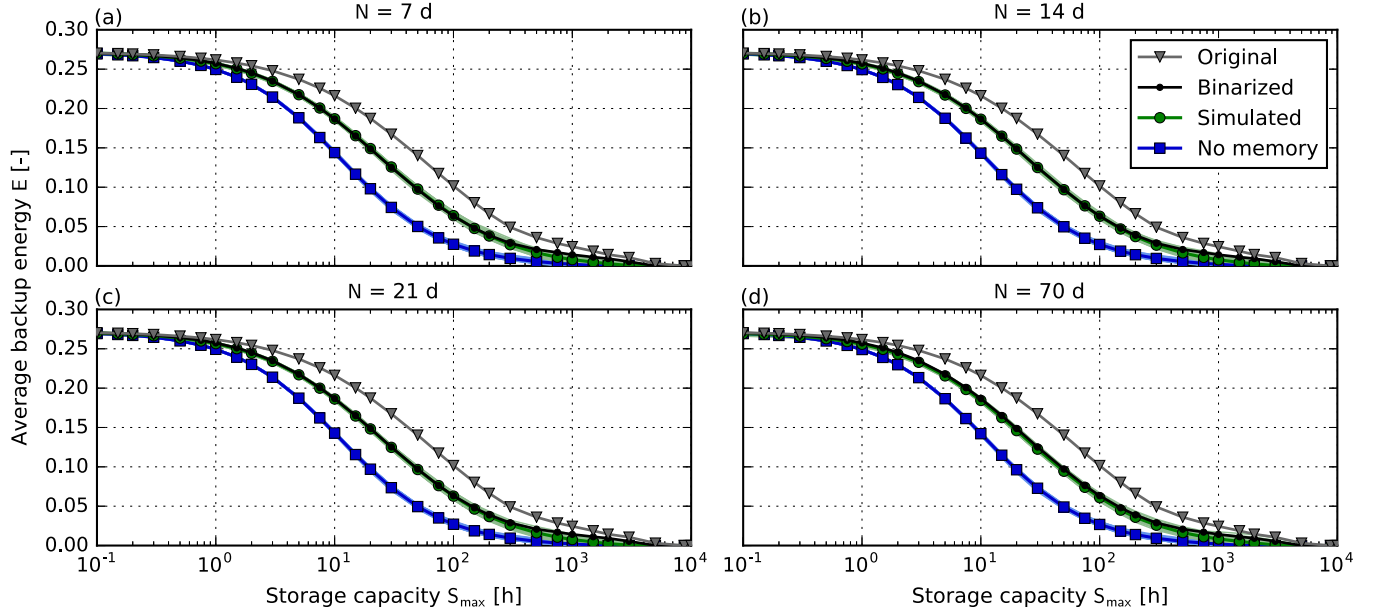
FIG. 4. Average backup energy E as a function of the storage capacity S_{max} for the original renewables.ninja [38] (gray triangles) and the binarized (black dots) time series and for an ensemble of 100 simulations using the additive binary Markov chain approach (green circles) for memory lengths of $N \in \{7, 14, 21, 70\}$ d and without memory, i.e., $F(r > 1) = 0$ (blue squares). The ensemble averages are represented as thick green and blue lines.

In order to keep the assumption of a fully renewable power system, the simulated data are scaled such that the relation $\langle R \rangle = \langle L \rangle = 1$ always holds.

The mapping of the original data to binary values reduces the backup need for all given S_{max} (gray triangles and black dots in Fig. 4) due to the loss of temporal correlation as discussed above. Thus, in order to evaluate our method, we compare the average backup energy of the simulated time series to that derived from the binarized time series (E_{bin}).

Considering the additive binary Markov chain model with memory (green lines in Fig. 4), we find that the average backup energy need $E_{bin}(S_{max})$ is reproduced with high accuracy for most storage capacities S_{max}. For $S_{max} > 100$ h, the backup energy is underestimated a little for all N. For high memory lengths, i.e., $N = 70$ d [Fig. 4(d)], the backup energy is slightly underestimated for all storage capacities S_{max}. This can be explained by the overall underestimation of the temporal correlation [cf. Fig. 3(d)].

The additive binary Markov chain without memory, i.e., $F(r > 1) = 0$, highly underestimates the backup need (blue lines in Fig. 4). This results from (i) the underestimation of long-range correlations (cf. Fig. 3) and (ii) the fact that the resting time distribution of underproduction and overproduction events cannot be reconstructed (see Fig. 5 below). This shows that the memory function $F(r)$ has to be considered for $r > 1$ h in order to get reasonable results for practically relevant characteristics of the time series.

3. Resting time distribution

A further benchmark for the performance of the additive binary Markov approach is to evaluate the distribution of the duration τ of periods for which $a(t) = 0$ (denoted by τ^-) and $a(t) = 1$ (denoted by τ^+), respectively [cf. Eqs. (19) and (20)].

A proper representation of this resting time distribution is particularly important for the backup and storage needs (cf. Fig. 4). This is especially true for the case of long-lasting scarcity events, i.e., $R(t) < 1$, which can deplete the storage and lead to high needs of backup energy. The resting time distribution of the synthetic data is derived from the ensemble of 100 simulations. Results are shown for $N = 14$ d in Fig. 5. We find that they hardly depend on the choice of N.

The distributions of the original and the simulated time series with memory are well comparable for a wide range of resting times (black and green lines in Fig. 5). Resting times for underproduction, i.e., $R \leqslant 1$, are slightly underestimated for $1 < \tau^- \leqslant 4$ h. Furthermore, resting times for overproduction, i.e., $R > 1$, tend to be overestimated for $\tau^+ \leqslant 3$ h and are underestimated a little for $3 < \tau^+ \leqslant 9$ h. This effect may result from the daily wind variation of the data. This leads to a maximum in the original resting time distribution at $\tau^{\pm} = 2$ h. This maximum cannot be represented by our model and is rather smoothed out.

Due to the diurnal variation in the wind power generation, we find local minima for resting times of about 12 and 24 h for both underproduction and overproduction. These local minima are well represented by our model.

The probability for a resting time $\tau^{\pm}$ to be longer than a certain value [shown as one minus the cumulative distribution function (CDF) in the insets in Fig. 5] decreases almost exponentially for the simulated data with memory. In contrast, for the original data the probabilities for long resting times decay more slowly. This most likely results from the small amount of original data showing high resting times and hence represents a finite-size effect. In the original data set, no resting times longer than $\tau_{max}^- = 719$ h and $\tau_{max}^+ = 305$ h are present such that it is impossible to compare probabilities in this regime.
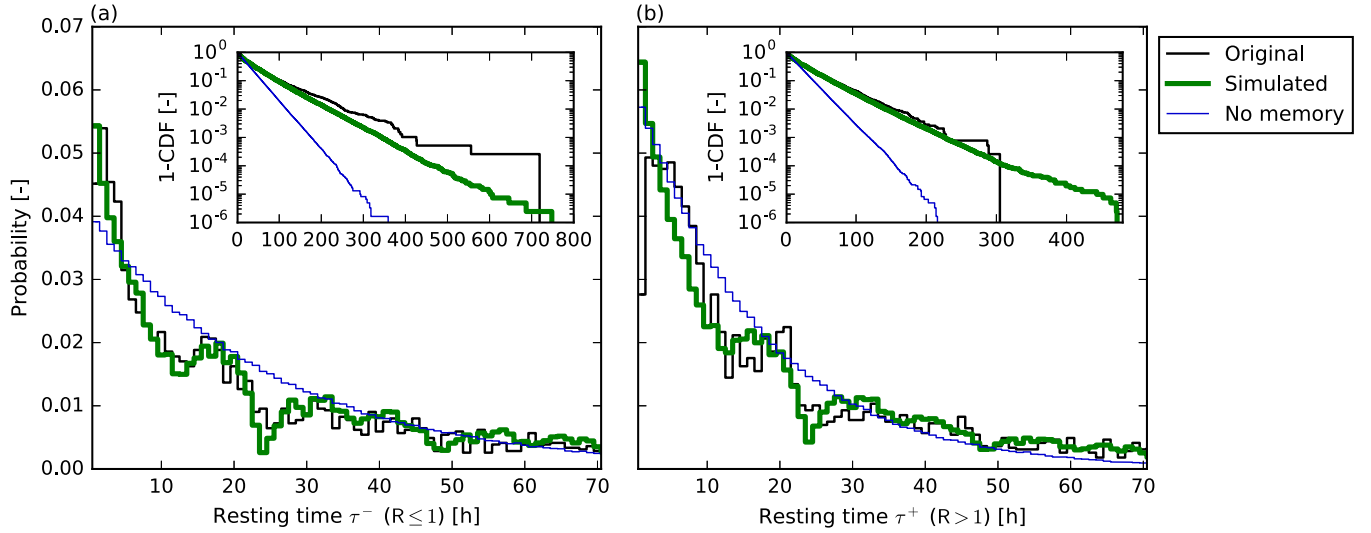
FIG. 5. Resting time distribution of the original renewables.ninja [38] time series (medium black lines) and the ensemble of 100 simulations using the additive binary Markov chain approach for a memory length of $N = 14$ d (thick green lines) and without memory, i.e., $F(r > 1) = 0$ (thin blue lines): (a) resting times for which the renewable generation R is constantly smaller than the (constant) load $L = 1$ (underproduction) and (b) resting times for overproduction. The probability of resting times is shown up to 70 h. The insets display the probability of long resting times using one minus the cumulative distribution function and a logarithmic y scale.

Without memory, i.e., $F(r > 1) = 0$ (blue lines in Fig. 5), the resting time distribution cannot be reconstructed properly. The curve also decreases exponentially, however, with the wrong rate. In particular, the probability for long durations is highly underestimated. This again shows that the higher terms of the memory function are essential for the stochastic model to work.

4. Conclusion

The additive binary Markov chain model reproduces the temporal characteristics of the binarized wind power time series to a high extent when using the full memory function $F(r)$. The autocorrelation function is reproduced almost perfectly within the memory time per construction, provided the memory time is not chosen unreasonably high. The backup and storage needs can be represented very well, especially for small and medium storage sizes. Finally, the resting time distribution of periods with below- and above-threshold wind generation can be reconstructed. Deviations are found for very large time spans, but the comparison to the original time series becomes tedious here due to finite-size effects. In comparison, an additive binary Markov model without memory clearly fails to reproduce characteristics essential for the power system operation and design. This proves that long memories must be taken into account in any stochastic model for wind power generation. The memory length N should be of the same order of magnitude as the time scales of the temporal correlations considered in the simulation (here $N = 7, \ldots, 21$ d).

C. Performance for different values of the renewable penetration

So far, we have assumed a fully renewable power system with 100% renewable generation on average, i.e., $\gamma = 1$. We now evaluate the model for different scenarios for the develop-

ment of renewable power sources quantified by the renewable penetration γ. This corresponds to different threshold values in the binarization of the empirical time series [cf. Eq. (18)]. We scale the wind power time series using (i) $\gamma = 0.5$ and (ii) $\gamma = 1.5$ [cf. Eq. (17)] such that the average wind generation equals (i) 50% and (ii) 150% of the load, indicating long resting times for underproduction and overproduction, respectively. As before, we consider an ensemble of 100 simulations and choose a memory length of $N = 14$ d.

1. Autocorrelation function

For both values of γ, the ensemble average of the autocorrelation function derived using the additive binary Markov model coincides almost perfectly with the autocorrelation function of the original binarized time series (see Fig. 6). Hence, we conclude that the scaling, and thus the choice of the threshold value, does not impact the representation of the temporal correlations in our model.

2. Backup and storage need

If the renewable penetration is chosen to be smaller than one, an average share of $1 - \gamma$ has to be provided by nonrenewable power plants. This share is deterministic. We are interested in the nondeterministic additional backup energy, which is given as $E_{\text{add}} = E - (1 - \gamma)$.

The (additional) average backup energy as a function of the storage capacity is shown in Fig. 7. A comparison with Fig. 4 shows that for $\gamma \neq 1$, the (additional) average backup energy is smaller than for $\gamma = 1$ for all storage capacities. This observation is well analyzed in the literature [11,12,24]. The average backup energy need can be reproduced almost perfectly for both a scaling with $\gamma = 0.5$ and a scaling with $\gamma = 1.5$. For $\gamma = 0.5$ and $S_{\text{max}} < 0.3$ h, E_{add} is slightly overestimated, whereas it is underestimated a little bit for
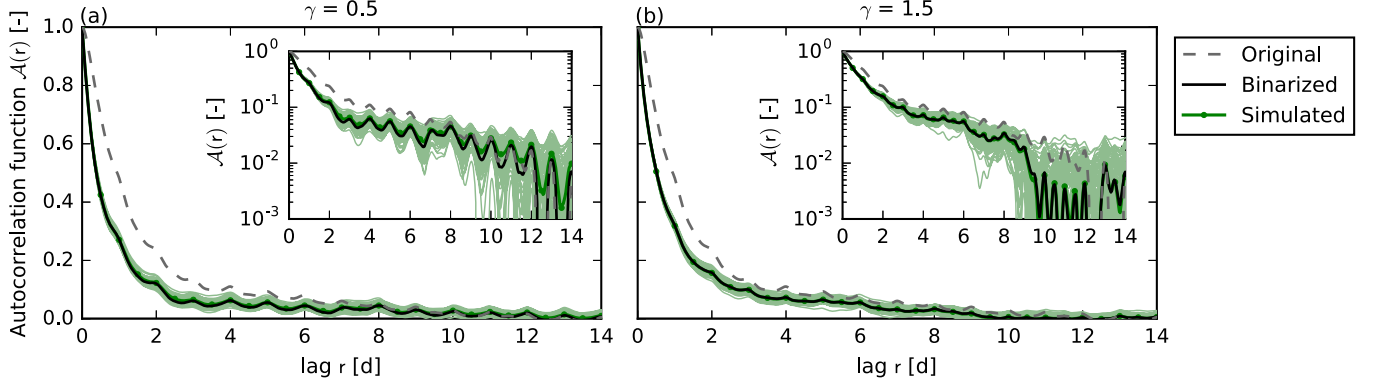
FIG. 6. Autocorrelation functions for the original renewables.ninja [38] (gray dashed lines) and the binarized (black solid lines) time series and an ensemble of 100 simulated time series using the additive binary Markov chain approach (green dots) for renewable penetrations of (a) $\gamma = 0.5$ and (b) $\gamma = 1.5$ and a memory length of $N = 14$ d. The ensemble average is represented as a thick green line.

$1.5 \leqslant S_{max} \leqslant 7.5$ h. This may be explained by the small differences in the resting time distribution (see Fig. 8 in the following section). For $\gamma = 1.5$, E is also slightly underestimated for $15 \leqslant S_{max} \leqslant 75$ h, which may be due to an underestimation of the probability for resting times τ^+ to be longer than about 200 h.

3. Resting time distribution

For $\gamma = 0.5$, the resting time distribution exhibits longer resting times τ^- for periods of underproduction and shorter resting times τ^+ for periods of overproduction [see Figs. 8(a) and 8(b)] compared to $\gamma = 1$ (cf. Fig. 5), per construction. In the case of $\gamma = 1.5$ the opposite is true [see Figs. 8(c) and 8(d)].

For resting times up to 70 h, the resting time distributions are represented well in most cases. For $\gamma = 0.5$, the model tends to underestimate the probability for $\tau^- \leqslant 6$ h [Fig. 8(a)]. Furthermore, as in the case of $\gamma = 1$, the model cannot reproduce the local maximum at $\tau^\pm \approx 2$–4 h very well. This is especially true for the case of overproduction, i.e., $R > 1$.

Considering the probability for resting times being longer than a certain value (insets in Fig. 8), the resting time distributions of the original and the simulated data are comparable and the slopes are almost identical. Due to the large ensemble of 100 simulations, the simulated data extend to longer resting times than the original data. For $\gamma = 1.5$ and τ^+, the probability for long durations is lower in the simulations than in the original time series.

4. Conclusion

In conclusion, the additive binary Markov model reconstructs well the temporal correlations and the backup energy as a function of the storage capacity for different scenarios of the development of the renewable expansion γ, i.e., for different threshold values. Furthermore, our model is also capable of representing very long resting times, especially in the case of underproduction, which is important for the sizing of the backup and storage infrastructure.
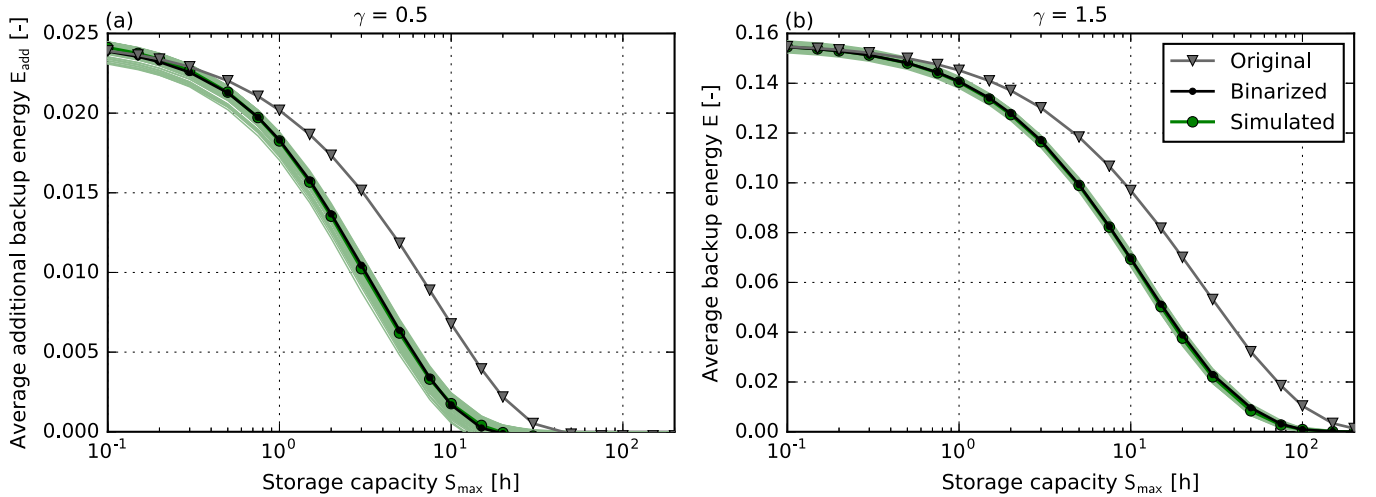

FIG. 7. Average (additional) backup need E (E_{add}) as a function of the storage capacity S_{max} for the original renewables.ninja [38] (gray triangles) and the binarized (black dots) time series and an ensemble of 100 simulated time series using the additive binary Markov chain approach (green circles) for renewable penetrations of (a) $\gamma = 0.5$ and (b) $\gamma = 1.5$ and a memory length of $N = 14$ d. The ensemble average is represented as a thick green line.
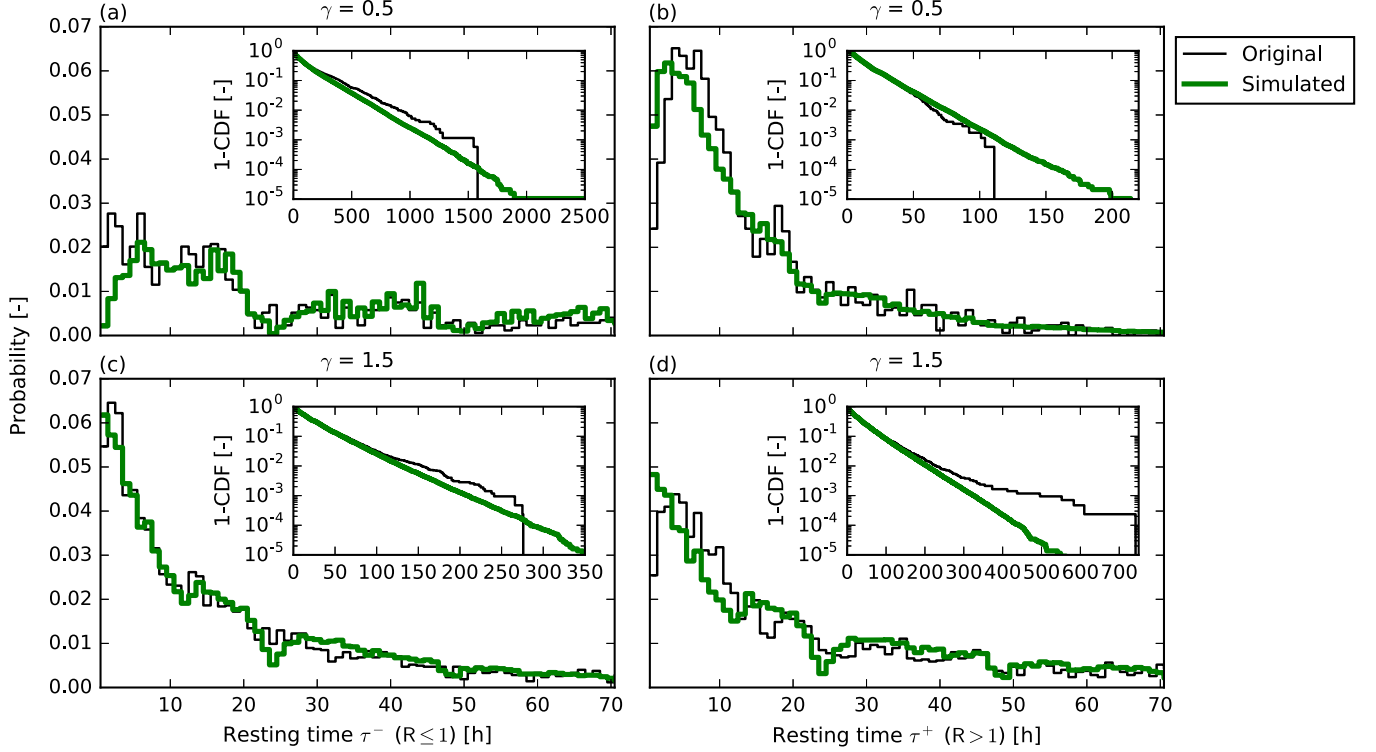
FIG. 8. Resting time distribution of the original renewables.ninja [38] time series (thin black lines) and the ensemble of 100 simulations using the additive binary Markov chain approach (thick green lines) for a memory length of $N = 14$ d. Resting times are shown for (a) and (c) underproduction ($R \leqslant 1$) and (b) and (d) overproduction ($R > 1$) for renewable penetrations of (a) and (b) $\gamma = 0.5$ and (c) and (d) $\gamma = 1.5$. The probability of resting times are shown up to 70 h. The insets display the probability of long resting times using one minus the cumulative distribution function and a logarithmic y scale.

V. EXTENSION OF THE MODEL TO NONBINARY WIND POWER DATA

One major drawback of our model is that it relies on binarized data, which underestimates the autocorrelation function as well as the average backup need E. We stress that this is not due to the stochastic modeling but solely due to the simplification of the empirical input data. We now present a method to overcome this deficiency while keeping the simplicity of the binary model.

A. Relation between duration and energy

The deterministic mapping (19) and (20) is replaced by a stochastic mapping

$$a(t) = 0 \Rightarrow R(t) = R^{-}, \tag{23}$$

$$a(t) = 1 \Rightarrow R(t) = R^{+}, \tag{24}$$

where R^{-} and R^{+} are random variables. They remain constant throughout any period of scarcity or oversupply, respectively, but depend on the duration of the period $\tau^{\pm}$.

The stochastic mapping is constructed from the empirical time series as follows: In each period of scarcity or oversupply, a certain amount of energy has to be provided by the backup and/or storage facilities or curtailed and/or stored in, respectively. The fundamental variable is thus the mismatch energy of one period. We use this energy as the basic variable for

the stochastic mapping and derive it as the area (A^{-} and A^{+}) under the curve for which $R(t) - \langle L \rangle < 0$ or $R(t) - \langle L \rangle > 0$, respectively (see Fig. 1),

$$A^{\pm}(\tau^{\pm}) = \int_{t_1}^{t_2} [R(t) - \langle L \rangle] dt, \tag{25}$$

with $\tau^{\pm} = t_2 - t_1$ and $\langle L \rangle = 1$.

A strong positive correlation exists between the duration $\tau^{\pm}$ of such a period and the corresponding mismatch energy $A^{\pm}$ (see Fig. 9, colored dots): The longer the duration of an event, the higher the mismatch energy, on average. We use this observation to assign mismatch energies to the periods of overproduction and underproduction simulated using the additive binary Markov chain. In order to do this, we derive a relation between $\tau^{\pm}$ and $A^{\pm}$ from the original renewables.ninja [38] data set. First, we define bins for τ^{+} and τ^{-} such that at least 15 values lie within each bin (cf. Fig. 9, horizontal lines). Here τ is a discrete variable. Thus, for short periods of overproduction and underproduction, for which more than 15 observations exist, the bin width is set to 1 h, the time resolution of the input data. The last bin is chosen such that it contains at least 10 observations. Second, we calculate the empirical cumulative distribution function (ECDF) of the corresponding energies A^{+} and A^{-} for every $\tau^{\pm}$ bin. Using a linear interpolation of the inverse of the ECDF, we can randomly assign energies $A^{\pm}$ for each duration $\tau^{\pm}$ encountered in the additive binary Markov model. If the
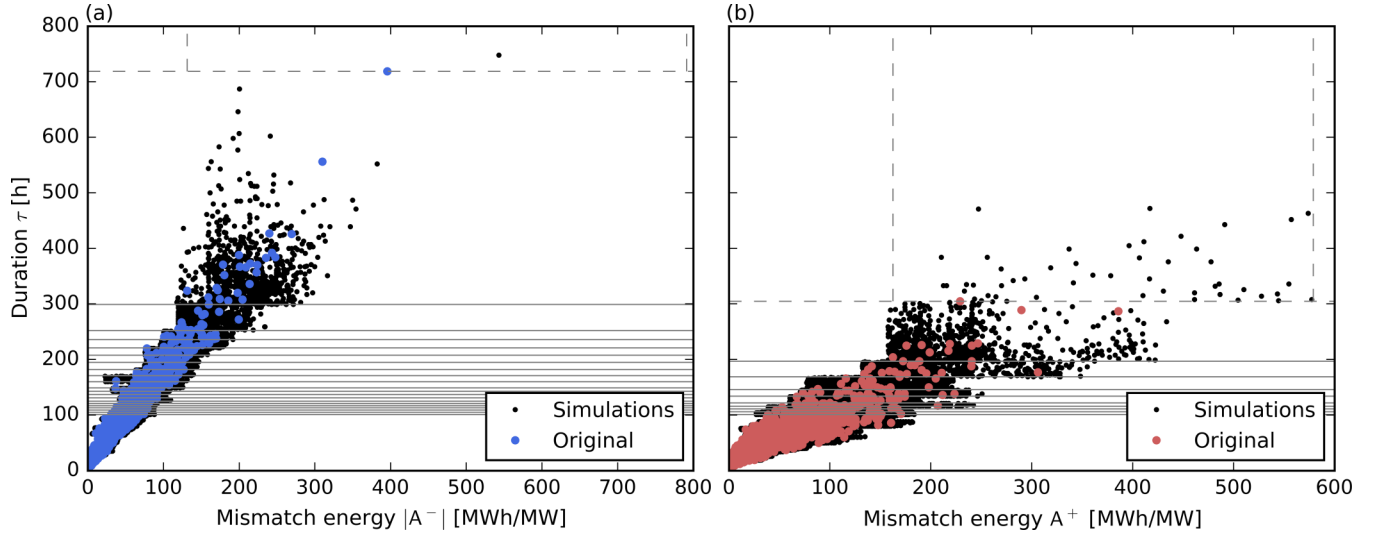
FIG. 9. Scatter plot of the duration τ and the corresponding mismatch energy of (a) underproduction ($|A^-|$) and (b) overproduction (A^+) for the original renewables.ninja [38] time series (large blue and red dots) and for an ensemble of 100 simulations using the additive binary Markov chain approach (small black dots) for a memory length of $N = 14$ d. The τ bins are chosen such that at least 15 of the original values lie within one bin (see the text). They are represented by the horizontal gray lines and only shown for $\tau \geqslant 100$ h. For each simulated τ, an energy value A is chosen according to the distribution of the original values within the bin. Energies which lie outside the range of the original values result from a rescaling as described in detail in the main text. If the simulated duration is higher than the maximum duration in the original time series (horizontal gray dashed line), the energy values are distributed uniformly within the region represented by the vertical gray dashed lines.

simulated duration is higher than any observed duration in the original time series, the mismatch energy is chosen randomly within an empirically chosen interval (vertical dashed lines in Fig. 9).

In a fully renewable power system, i.e., $\gamma = 1$, the sum of all mismatch energies has to be zero, i.e., $\sum A^+ + \sum A^- = 0$. In the simulations, this sum is usually only close to zero. Thus, the energies have to be rescaled. If $\sum A^+ < \sum |A^-|$, the A^+

are scaled to higher values

$$A_{\text{new}}^+ = \frac{\sum |A^-|}{\sum A^+} A^+ \qquad (26)$$

and vice versa. It should be noted that this scaling can be done in several ways. We chose to scale to higher absolute values because we find that the ensemble mean of the average mismatch energy ($\frac{1}{100} \sum_{n=1}^{100} \langle A^{\pm} \rangle$) is smaller than the average
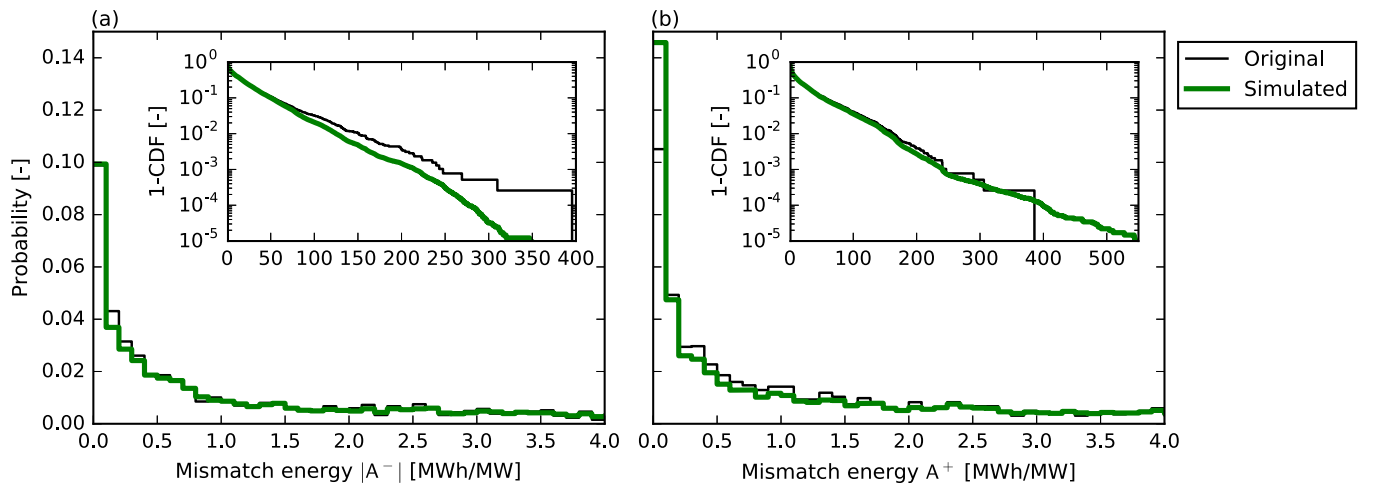

FIG. 10. Marginal distribution of the energies corresponding to (a) underproduction ($|A^-|$) and (b) overproduction (A^+) for the original renewables.ninja [38] time series (thin black lines) and the ensemble of 100 simulations using the additive binary Markov chain approach for a memory length of $N = 14$ d (thick green lines). Probabilities are shown for small mismatch energies. The insets focus on large mismatch energies using one minus the cumulative distribution function and a logarithmic y scale.
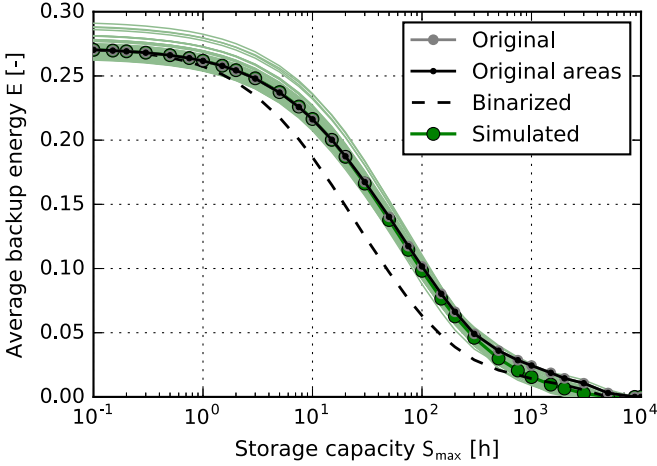
FIG. 11. Average backup energy E as a function of the storage capacity S_{max} for the ensemble of 100 simulations using the additive binary Markov chain approach combined with the mismatch energy distribution (green circles) for a memory length of $N = 14$ d. The ensemble average is represented as a thick green line. Results have to be compared to those of the piecewise averaged original renewables.ninja [38] time series (black dots). Additionally, results for the original time series (gray circles) and for the binarized time series (black dashed line) are shown.
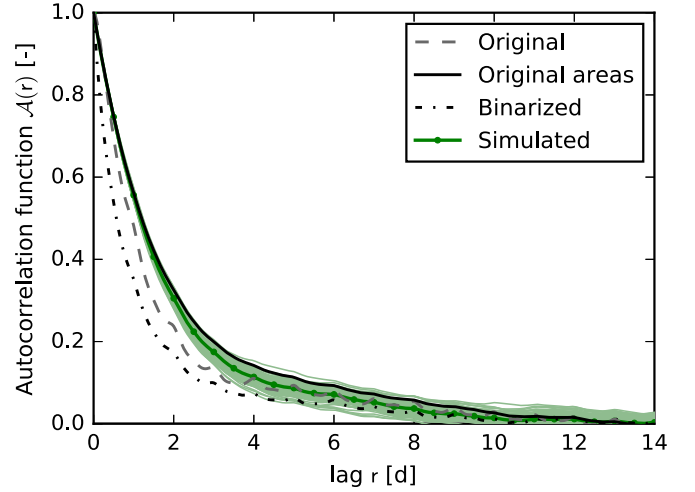

FIG. 12. Autocorrelation function for the ensemble of 100 simulations using the additive binary Markov chain approach combined with the mismatch energy distribution (green dots) for a memory length of $N = 14$ d. The ensemble average is represented as thick green line. Results have to be compared to those of the piecewise averaged original renewables.ninja [38] time series (black solid line). Additionally, results for the original time series (gray dashed line) and for the binarized time series (black dashed line) are shown.

mismatch energy of the original time series. The black dots in Fig. 9 resemble the resulting τ-A distribution of the ensemble of 100 simulations of the additive Markov chain for $N = 14$ d and $\gamma = 1$.

B. Evaluation of the method

1. Marginal distributions of the mismatch energies

The marginal distributions of A^+ and A^- are almost identical for the original and the simulated time series for a wide range of values (Fig. 10). The simulations slightly underestimate the probabilities for $0.1 < |A^-| < 0.5$ and $0.1 < A^+ < 1.2$. This results directly from the deviations in the resting time distribution (cf. Fig. 5): As τ and A are strongly correlated, an underestimated probability for small values of τ most certainly leads to an underestimated probability for small mismatch energies as well. For high absolute values of A^+ and A^- (insets of Fig. 10), the same arguments hold: The probabilities for $|A^-| > 50$ and $150 < A^+ < 386$ are smaller in the simulations compared to the original time series due to the lower likelihood for durations in the range of $100 < \tau^- < 719$ and $100 < \tau^+ < 305$, respectively, which results from finite-size effects.

2. Backup and storage need

To test our model, we again derive the average backup energy E as a function of the storage capacity S_{max}. In order to apply Eqs. (10) and (11), we calculate the piecewise averaged wind generation time series from the mismatch energies as follows:

$$R^{\pm}(t) = \langle L \rangle + \frac{A^{\pm}(\tau^{\pm})}{\tau^{\pm}}. \tag{27}$$

The average over the ensemble of 100 simulations almost perfectly coincides with the average backup energy E of the original time series for $S_{max} < 50$ h (see Fig. 11). For higher storage capacities, E is slightly underestimated as already observed before [cf. Fig. 4(b)]. The spread of the ensemble, however, is higher because the mismatch energies are not perfectly matched.

3. Autocorrelation function

The autocorrelation of the ensemble of the simulated time series is shown in Fig. 12. As we are dealing with piecewise constant positive and negative values $A^+(\tau^+)$ and $A^-(\tau^-)$, respectively, we have to compare the autocorrelation functions of the simulations to the autocorrelation function derived from the A^+ and A^- values of the original time series (labeled "Original areas"). We find that the model underestimates the autocorrelation function for all time lags. One reason for this could be that the assignment of mismatch energies is done in a too simplified way. Another reason for the underestimation could be that we tend to underestimate the average mismatch energies, even after the rescaling [cf. Eq. (26)]. Thus, more information on the dependence of τ and A should be taken into account in order to reconstruct the autocorrelation function more exactly.

VI. CONCLUSION

Temporal correlations of wind power generation are a crucial factor to be considered for the integration of high shares of renewables [4,5,5,6,11–15,17–19]. However, it is difficult to model these correlations and simple Markov models often fail to represent them properly [20,26,29].

We introduced a model to accurately reproduce the temporal correlations of wind power generation time series, focusing on

the resting time of binary underproduction and overproduction events. For this purpose, we used the concept of additive binary Markov chains introduced in [31,33]. This concept allowed us to model binary wind generation data by employing the empirical autocorrelation function of a binarized input time series.

For a wide range of memory lengths, the autocorrelation function, the average backup need as a function of the storage capacity, and the resting time distribution of the binarized input time series can be reproduced almost perfectly by the additive binary Markov model. A short memory length better represents short-range correlations, whereas a long memory length leads to a better representation of the long-range correlations. Furthermore, we showed that the exact choice of the renewable penetration γ hardly impacts the results.

In order to transform the simulated binary time series back to a nonbinary time series, we used the joint probability of the resting time τ and the mismatch energy A. The higher the resting time, the higher the mismatch energy, on average. The application of this relation allowed us to derive the average backup energy as a function of the storage capacity and to compare it directly to that of the original time series. We found that our model is able to reconstruct this function for storage capacities of up to about 100 h. The autocorrelation function is

underestimated, potentially indicating that the relation between τ and A is not sufficiently well captured.

The presented model is fairly simple to implement and still captures essential features of wind generation time series, especially long memory times. It thus overcomes essential problems of previous approaches [20,26–30]. In the present paper we demonstrated the ability of the additive binary Markov model for wind time series on national scales. However, it may also be used to capture temporal characteristics in different regimes such as smaller spatial and/or temporal scales. For future applications, this model can be used to learn more about the temporal structure of high- and low-wind periods, for example, when planning to combine a wind farm with a storage device.

ACKNOWLEDGMENTS

[1] UNFCCC, Adoption of the Paris Agreement, Report No. FCCC/CP/2015/L.9/Rev.1, http://unfccc.int/resource/docs/2015/cop21/eng/l09r01.pdf (2015).

[2] J. Rogelj, G. Luderer, R. C. Pietzcker, E. Kriegler, M. Schaeffer, V. Krey, and K. Riahi, Nat. Clim. Change **5**, 519 (2015).

[3] J. Rogelj, M. Den Elzen, N. Höhne, T. Fransen, H. Fekete, H. Winkler, R. Schaeffer, F. Sha, K. Riahi, and M. Meinshausen, Nature (London) **534**, 631 (2016).

[4] R. Sims *et al.*, in *IPCC Special Report on Renewable Energy Sources and Climate Change Mitigation*, edited by O. Edenhofer *et al.* (Cambridge University Press, Cambridge, 2011).

[5] H. Bloomfield, D. J. Brayshaw, L. C. Shaffrey, P. J. Coker, and H. Thornton, Environ. Res. Lett. **11**, 124025 (2016).

[6] J. Olauson, M. N. Ayob, M. Bergkvist, N. Carpman, V. Castellucci, A. Goude, D. Lingfors, R. Waters, and J. Widén, Nat. Energy **1**, 16175 (2016).

[7] C. Schiel, P. G. Lind, and P. Maass, Sci. Rep. **7**, 11562 (2017).

[8] B. Schäfer, M. Matthiae, X. Zhang, M. Rohden, M. Timme, and D. Witthaut, Phys. Rev. E **95**, 060203 (2017).

[9] K. Schmietendorf, J. Peinke, and O. Kamps, Eur. Phys. J. B **90**, 222 (2017).

[10] B. Schäfer, C. Beck, K. Aihara, D. Witthaut, and M. Timme, Nat. Energy **3**, 119 (2018).

[11] D. Heide, L. von Bremen, M. Greiner, C. Hoffmann, M. Speckmann, and S. Bofinger, Renew. Energy **35**, 2483 (2010).

[12] M. G. Rasmussen, G. B. Andresen, and M. Greiner, Energy Policy **51**, 642 (2012).

[13] F. Díaz-González, A. Sumper, O. Gomis-Bellmunt, and R. Villafáfila-Robles, Renew. Sustain. Energy Rev. **16**, 2154 (2012).

[14] P. Elsner, M. Fischedick, and D. U. Sauer, *Flexibilitätskonzepte für die Stromversorgung 2050* (Acatech, Muenchen, 2015).

[15] P. Milan, M. Wächter, and J. Peinke, Phys. Rev. Lett. **110**, 138701 (2013).

[16] M. Anvari, G. Lohmann, M. Wächter, P. Milan, E. Lorenz, D. Heinemann, M. R. R. Tabar, and J. Peinke, New J. Phys. **18**, 063027 (2016).

[17] D. Schlachtberger, S. Becker, S. Schramm, and M. Greiner, Energy Convers. Manage. **125**, 336 (2016).

[18] D. Cannon, D. Brayshaw, J. Methven, P. Coker, and D. Lenaghan, Renew. Energy **75**, 767 (2015).

[19] J. Weber, J. Wohland, M. Reyers, J. Moemken, C. Hoppe, J. G. Pinto, and D. Witthaut, arXiv:1711.05569.

[20] T. Pesch, S. Schröders, H. J. Allelein, and J. F. Hake, New J. Phys. **17**, 055001 (2015).

[21] C. M. Grams, R. Beerli, S. Pfenninger, I. Staffell, and H. Wernli, Nat. Clim. Change **7**, 557 (2017).

[22] J. Wohland, M. Reyers, J. Weber, and D. Witthaut, Earth System Dyn. **8**, 1047 (2017).

[23] M. M. Bandi, Phys. Rev. Lett. **118**, 028301 (2017).

[24] T. V. Jensen and M. Greiner, Eur. Phys. J. Spec. Top. **223**, 2475 (2014).

[25] I. Staffell and S. Pfenninger, Energy **145**, 65 (2018).

[26] K. Brokish and J. Kirtley, *Proceedings of the Power Systems Conference and Exposition, 2009 (PSCE'09)* (IEEE, Piscataway, 2009), pp. 1–6.

[27] G. Papaefthymiou and B. Klockl, IEEE Trans. Energy Convers. **23**, 234 (2008).

[28] H. V. Haghi, M. T. Bina, and M. A. Golkar, IEEE Trans. Sustain. Energy **4**, 838 (2013).

[29] R. Carapellucci and L. Giordano, Energy **49**, 298 (2013).

[30] R. Carapellucci and L. Giordano, Appl. Energy **101**, 541 (2013).

[31] O. V. Usatenko, V. A. Yampol'skii, K. E. Kechedzhy, and S. S. Mel'nyk, Phys. Rev. E **68**, 061107 (2003).

[32] O. V. Usatenko and V. A. Yampol'skii, Phys. Rev. Lett. **90**, 110601 (2003).

[33] S. S. Melnyk, O. V. Usatenko, and V. A. Yampol'skii, Physica A **361**, 405 (2006).

[34] S. S. Melnyk, O. V. Usatenko, V. A. Yampol'skii, S. S. Apostolov, and Z. A. Maiselis, J. Phys. A: Math. Gen. **39**, 14289 (2006).

[35] N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry* (Elsevier, Amsterdam, 1992), Vol. 1.

[36] C. W. Gardiner, *Handbook of Stochastic Methods* (Springer, Berlin, 1985), Vol. 3.

[37] R. A. Rodriguez, S. Becker, G. B. Andresen, D. Heide, and M. Greiner, Renew. Energy **63**, 467 (2014).

[38] I. Staffell and S. Pfenninger, Energy **114**, 1224 (2016).

[39] M. M. Rienecker, M. J. Suarez, R. Gelaro, R. Todling, J. Bacmeister, E. Liu, M. G. Bosilovich, S. D. Schubert, L. Takacs, G.-K. Kim *et al.*, J. Clim. **24**, 3624 (2011).

[40] M. Reyers, J. Moemken, and J. G. Pinto, Int. J. Climatol. **36**, 783 (2016).

[41] J. Moemken, M. Reyers, H. Feldmann, and J. G. Pinto (unpublished).

[42] G. C. Thomann and M. J. Barfield, IEEE Trans. Energy Convers. **3**, 44 (1988).