Downscaling near-surface atmospheric fields with multi-objective Genetic Programming

Tanja Zerenner¹, Victor Venema, Petra Friederichs, Clemens Simmer

Meteorological Institute, University of Bonn, Auf dem Huckel 16, 53121 Bonn and Meckenheimer Allee 176, 53115 Bonn, Germany

A R T I C L E   I N F O

Article history:
Received 17 December 2015
Received in revised form 27 April 2016
Accepted 11 June 2016
Available online 5 July 2016

Keywords:
Statistical downscaling
Disaggregation
Evolutionary computation
Machine learning
Pareto optimality
Coupled modeling

A B S T R A C T

We present a new Genetic Programming based method to derive downscaling rules (i.e., functions or short programs) generating realistic high-resolution fields of atmospheric state variables near the surface given coarser-scale atmospheric information and high-resolution information on land surface properties. Such downscaling rules can be applied in coupled subsurface-land surface-atmosphere simulations or to generate high-resolution atmospheric input data for offline applications of land surface and subsurface models. Multiple features of the high-resolution fields, such as the spatial distribution of subgrid-scale variance, serve as objectives. The downscaling rules take an interpretable form and contain on average about 5 mathematical operations. The method is applied to downscale 10 m-temperature fields from 2.8 km to 400 m grid resolution. A large part of the spatial variability is reproduced, also in stable nighttime situations, which generate very heterogeneous near-surface temperature fields in regions with distinct topography.

© 2016 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

With growing computational power integrated modeling platforms coupling atmospheric, land surface and subsurface models are increasingly used to account for interactions and feedbacks between the different components (e.g., Shrestha et al., 2014). The feedback processes are largely driven by the turbulent exchange fluxes of energy, moisture and momentum at the interface between land surface and atmosphere. The use of spatially averaged parameters or state variables at the land surface or the lower atmospheric boundary layer (ABL) can introduce biases in the flux estimation. In current atmospheric models for numerical weather prediction, which are typically applied at scales of few kilometers, heterogeneities at smaller scales are mostly neglected.

Subgrid-scale parameterization of the land surface like tile, mosaic or mixture approaches significantly improve the estimation of the surface fluxes (e.g., Avissar and Pielke, 1989; Koster and Suarez, 1992; Leung and Ghan, 1995; Schlünzen and Katzfey, 2003). Shao et al. (2001) showed that also the representation of the subgrid-scale atmospheric heterogeneity improves the flux estimates.

The explicit subgrid approach by Seth et al. (1994) allows to combine the subgrid representation of the land surface with downscaled atmospheric forcings. In the explicit approach each atmospheric model grid box covers \(N \times N \) land surface columns, i.e., a higher resolution land surface scheme is nested into a coarser resolution atmospheric model (see also Giorgi et al. (2003) and Ament and Simmer (2006) for discussion). This is analogue to coupling a coarser atmospheric model with a high-resolution land surface model, as it is often done in the aforementioned integrated modeling platforms. This approach is feasible because of the comparatively low computational cost of land surface and subsurface models.

Besides the potential to improve the estimation of the turbulent exchange fluxes, downscaling of the near-surface atmospheric state variables can provide better forcing data for land surface, subsurface and agricultural models. This is important as besides the turbulent exchange coefficients also many processes at the earth's surface, e.g., related to vegetation, are nonlinear. Furthermore, the representation of runoff production or snow melt, which are threshold dependent, would benefit from taking subgrid atmospheric variability into account.

Seth et al. (1994) introduced a simple atmospheric downscaling for the global climate scale (from 3.0° to 0.5° ≈ 50 km), which for instance corrects near-surface temperature using the model simulated ground temperature or topographic height at the high resolution. Fiddes and Gruber (2014) presented a more advanced...
physically based downscaling scheme, TopoSCALE, to create high-resolution forcing data for land surface models from global climate reanalysis (from 0.75° to ≤100 m), using fine-scale topography information from a high-resolution digital elevation model. A similar approach was taken by Schomburg et al. (2010, 2012), who developed a downscaling scheme at the mesoscale (from 2.8 km to 400 m) by statistically evaluating high-resolution atmospheric model runs. The scheme leads to improvements for certain variables (e.g., near-surface pressure) and weather conditions (e.g., near-surface temperature in unstable atmospheres). As processes in the lower ABL can be complex and highly nonlinear, the conditional linear regression approach used in Schomburg et al. (2010) appears not to be sufficient to capture many of the processes acting in the lower ABL.

In this study we introduce a more flexible approach to detect relations (downscaling rules) that generate high-resolution atmospheric fields from coarse atmospheric information and high-resolution information on land surface characteristics. We employ Genetic Programming (GP), a machine learning method from the area of evolutionary computation (e.g., Koza, 1992; Banzhaf et al., 1997). Like artificial neural networks GP allows to flexibly model complex nonlinear and multivariate relations with the advantage that the downscaling rules take the form of equations or program code, which is readable, and thus can be checked for physical consistency.

Coulibaly (2004), Liu et al. (2008) and Hashmi et al. (2011) employed GP based methods to downscale temperature and/or precipitation from global climate model output to a station or catchment mean. The results were compared to the Statistical Down-Scaling Model (SDSM) by Wilby et al. (2002). In all three studies the GP based methodologies performed better than the SDSM. In Coulibaly (2004) and Hashmi et al. (2011) the downscaling models resulting from GP and SDSM were explicitly compared showing that the GP model not only performed better, but also required less predictor variables. Liu et al. (2008) additionally compared the GP results against a feed forward neural net. Both methods performed about equally well.

Unlike previous studies which employ GP for atmospheric downscaling, we aim at the downscaling of coherent spatial fields. To this goal we employ a multi-objective approach, because a regression aiming solely at the minimization of the root mean square error (RMSE) is known to have too small variance (e.g., Hastie et al., 2009). Instead of aiming at predicting the expected value, we aim at downscaling the expected value of the temperature anomalies given surface characteristics and coarse atmospheric state. Such an estimator is known to have too small variance (e.g., Hastie et al., 2009). Instead of aiming at predicting the expected value, we aim at downscaling rules returning realizations from an unknown multivariate probability density function (PDF). We do not optimize solely the RMSE, but also objectives that quantify the spatial variance on the subgrid-scale and the cumulative density functions (CDFs) of the full fields.

When we formulate the downscaling problem as a multi-objective optimization problem, we face, however, the following problems. Minimizing the sum of different objectives is problematic, since they may have different units and ranges. Even with an appropriate scaling procedure there is a risk of treating the objectives unequally or getting trapped in a local minimum. Firstly, we can never know, what is the minimum value of each objective that can be achieved by the regression. Thus, designing an appropriate scaling procedure is difficult and one would need to decide on the relative importance of the different objectives in advance. Secondly, adding multiple, conflicting objectives very likely results in a fitness function with multiple local minima, which makes optimization more difficult. To avoid these problems, we have implemented fitness calculation according to the Strength Pareto Evolutionary Algorithm (SPEA) by Zitzler and Thiele (1999), instead of using a single (weighted) fitness or cost function. Approaches for multi-objective optimization like SPEA are widely used in evolutionary computation. In SPEA the fitness calculation during the fitting procedure is based on an intercomparison of the different models.

3. Methods

We downscale near-surface atmospheric fields by establishing a statistical relation (downscaling rule) between the coarse atmospheric model output and the high-resolution atmospheric fields using quasi-static high-resolution land surface information. Thus, we assume that the structure of the atmospheric boundary layer near the surface is significantly influenced by land surface heterogeneity.

A rule search algorithm based on Genetic Programming is set up, which can potentially detect multivariate and nonlinear downscaling rules. Such rules are much less complex than running the full 3D-model at high resolution. It is not expected that the downscaling rules reproduce the exact high-resolution references. Due to turbulence for instance, there will always be a remaining component of the fine-scale fields that cannot be reconstructed.

We take a multi-objective approach that allows multiple characteristics of the fine-scale fields to be incorporated during the fitting of the regression model. Minimizing only the root mean square error (RMSE) would result in downscaling rules predicting the expected value of the temperature anomalies given surface characteristics and coarse atmospheric state. Such an estimator is known to have too small variance (e.g., Hastie et al., 2009). Instead of aiming at predicting the expected value, we aim at downscaling rules returning realizations from an unknown multivariate probability density function (PDF). We do not optimize solely the RMSE, but also objectives that quantify the spatial variance on the subgrid-scale and the cumulative density functions (CDFs) of the full fields.

When we formulate the downscaling problem as a multi-objective optimization problem, we face, however, the following problems. Minimizing the sum of different objectives is problematic, since they may have different units and ranges. Even with an appropriate scaling procedure there is a risk of treating the objectives unequally or getting trapped in a local minimum. Firstly, we can never know, what is the minimum value of each objective that can be achieved by the regression. Thus, designing an appropriate scaling procedure is difficult and one would need to decide on the relative importance of the different objectives in advance. Secondly, adding multiple, conflicting objectives very likely results in a fitness function with multiple local minima, which makes optimization more difficult. To avoid these problems, we have implemented fitness calculation according to the Strength Pareto Evolutionary Algorithm (SPEA) by Zitzler and Thiele (1999), instead of using a single (weighted) fitness or cost function. Approaches for multi-objective optimization like SPEA are widely used in evolutionary computation. In SPEA the fitness calculation during the fitting procedure is based on an intercomparison of the different models.
Further, a finite set of so called Pareto optimal models (downscaling rules) is returned. In the following we first give a summary of the 3-step downscaling scheme by Schomburg et al. (2010, 2012), which we will later use for comparison. Then the multiple objectives and the concept of Pareto optimality are introduced. Finally, a general introduction to GP is given, and the multi-objective GP is explained step by step. Large parts of the code are based on the GPLAB package for Matlab by Silva and Almeida (2003).


The downscaling scheme by Schomburg et al. (2010, 2012) consists of 3 steps, which are consecutively applied. In step 1 a bi-quadratic spline interpolation is used to smooth the coarse field

\[
y_{ij} = y(i,j) = a_1 + a_2 i + a_3 j + a_4 i^2 + a_5 j^2
\]

with \( y \) denoting an atmospheric variable, for instance temperature, \((i,j)\) the grid point indices on the fine scale, and \( a_1,...,a_5 \) the regression coefficients. To estimate the regression coefficients, five constraints are introduced: The derivatives of Eq. (1) at the four edges of the coarse pixel are required to equal the gradient between the coarse pixel and the corresponding neighboring pixel, and the coarse pixel mean is conserved.

In step 2 deterministic rules are employed to reduce the differences between the spline-interpolated and the high-resolution reference field. The deterministic rules are based on a linear regression model with high-resolution data of surface properties, such as references between the spline-interpolated and the high-resolution field. The deterministic rules are based on a linear regression and land surface fields and high-resolution surface properties for training data subsets defined by indicators, e.g., for data with a temperature gradient of the lowest 25 m.

\[ T^* = \text{if } T_{g25} < 0.0058 \text{ km}^{-1} \]
\[ \text{then } -0.0084 \text{ km}^{-1} \times hsurfa \]
\[ \text{else } 0, \]

i.e., if the temperature gradient is lower than 0.0058 km\(^{-1}\), the temperature anomaly is a linear function of the topographic height anomaly \( hsurfa \).

In the optional step 3 the small-scale variability that has not been reproduced by step 1 and 2 is added as temporally autoregressive Gaussian noise. The strength of the noise within each coarse pixel is estimated from the variance between neighboring coarse pixels. Spatial dependency on the sub pixel scale is neglected.

3.2. Objectives

The first step to set up the multi-objective Genetic Programming is the definition of the objectives. As in the 3-step approach by Schomburg et al. (2010, 2012) the downscaling rules are aimed at predicting anomalies, i.e., the differences between spline-interpolated fields and high-resolution reference. In the following this is not always explicitly mentioned. Note further, that the coarse pixel mean is conserved by the downscaling to avoid biases. This is done by subtracting the mean anomaly predicted over a coarse pixel from the predicted anomalies at every pixel on the fine scale.

The first three objectives, a neighborhood based root mean square error, the error of the spatial distribution of variance and the error of the discrete PDFs, quantify the quality of the downscaling rules. The complexity of a downscaling rule serves as fourth objective. In our case, each objective \( s_i \) is negatively oriented in the sense that it decreases the better the fit. The motivation and precise definition of the four objectives are given in the following.

Our first objective is the root mean square error (RMSE) between predicted \( y^D \) and observed anomalies \( y^O \) at each grid point.

\[
\text{RMSE} = \sqrt{\frac{1}{nti nji} \sum_{i,j} \sum_{k \in U(i,j)} \min (y^O_{ijkl} - y^D_{ijkl})^2}
\]

(2)

where neighborhood \( U(i,j) = \{(i,j), (i,j+1), (i,j-1), (i+1,j), (i-1,j)\} \) contains the four direct neighbors of each grid point \( (i,j) \) (Fig. 1). Here, \( n_t \) and \( n_i \) denote the total number of pixels in \( x \)- and \( y \)-direction and \( n_t \) the number of fields (Fig. 1).

In order to reconstruct the subgrid-scale spatial variability, the mean error of the subgrid-scale standard deviation (ME(STD)) serves as second objective. We define the ME(STD) as follows. Let \( V(p,q) \) denote a pixel on the coarse scale containing \( 7 \times 7 \) pixels on the finer scale (Fig. 1). Unlike \( U(i,j) \), \( V(p,q) \) is not defined as a sliding neighborhood, but via a fixed grid, i.e., the grid of the coarse scale model output. The coarse pixel standard deviation \( \sigma(y^R_{p,q}) \) of the reference field \( t \) reads

\[
\sigma(y^R_{p,q}) = \sqrt{\frac{1}{7 \times 7} \sum_{i,j \in V(p,q)} (y^R_{ij} - y^R_{p,q})^2},
\]

(3)

with \( y^R_{p,q} \) denoting the coarse pixel mean. Again \( y \) represents the anomalies. As the spline-interpolation conserves the coarse pixel

Table 1
Simulation dates and weather situations used for training and validation of the GP runs. The right column lists the time steps we have extracted from the full data set from Schomburg et al. (2010) to reduce computational cost.

<table>
<thead>
<tr>
<th>Date</th>
<th>Weather</th>
<th>Time steps [UTC]</th>
</tr>
</thead>
<tbody>
<tr>
<td>27 Aug. 2007</td>
<td>Varying cloud cover, no precipitation</td>
<td>03:00–04:00, 15:00–16:00</td>
</tr>
<tr>
<td>14 Oct. 2007</td>
<td>Clear sky</td>
<td>11:00–12:00, 23:00–24:00</td>
</tr>
<tr>
<td>10 Mar. 2008</td>
<td>Strong winds, variable clouds and precipitation</td>
<td>10:00–11:00, 22:00–23:00</td>
</tr>
<tr>
<td>2 May 2008</td>
<td>Clouds and precipitation</td>
<td>00:00–01:00, 12:00–13:00</td>
</tr>
<tr>
<td>10 May 2008</td>
<td>Clear sky</td>
<td>01:00–02:00, 13:00–14:00</td>
</tr>
<tr>
<td>7 June 2008</td>
<td>Convective clouds and precipitation</td>
<td>05:00–06:00, 17:00–18:00</td>
</tr>
<tr>
<td>21 July 2008</td>
<td>Synoptically driven stratiform precipitation</td>
<td>09:00–10:00, 21:00–22:00</td>
</tr>
<tr>
<td>28 Aug. 2008</td>
<td>Cloudy, some precipitation</td>
<td>07:00–08:00, 19:00–20:00</td>
</tr>
</tbody>
</table>

T. Zerenner et al. / Environmental Modelling & Software 84 (2016) 85–98
mean, the coarse pixel mean anomaly (high-resolution minus spline-interpolated field) equals zero, \( \bar{y}_{jq}^{p} = \bar{y}_{jq}^{s} = 0 \). We can now define the second objective as

\[
ME(\text{STD}) = \frac{7 \times 7}{n_{f} n_{t}} \sum_{t=1}^{n_{t}} \sigma(y_{jq}^{p}) - \sigma(y_{jq}^{g}).
\]  

(4)

One possibility to make two fields look similar is to make their values similar in distribution. Our third objective is to minimize the differences between the PDFs of the downscaled and the reference field. This objective considers the full fields and ignores the location of the particular values. The difference is quantified by the integrated quadratic distance (IQD) between the two cumulative distribution functions (CDFs) \( F \) and \( G \) (Thorarinsdottir et al., 2013):

\[
IQD = \int_{-\infty}^{\infty} (F(x) - G(x))^2 \, dx.
\]

(5)

We apply the IQD to the CDFs of the discretized temperature distribution with a bin width of 0.25 K. Let \( H(y_{jq}^{f}) \) be the discretized CDF of the high-resolution reference field at time step \( t \), and accordingly, \( H(y_{jq}^{s}(t)) \) the CDF of the corresponding downscaled field. A single bin is denoted by \( H_{t} \). From the CDFs we calculate the IQD for each field (time step) separately (Fig. 1) and take the mean over all time steps (i.e., the complete training data set) as objective:

\[
IQD = 1/n_{t} \sum_{t=1}^{n_{t}} \sum_{1}^{n_{j}} \left( H_{t}(y_{jq}^{f}) - H_{t}(y_{jq}^{s}) \right)^{2},
\]

(6)

with \( n_{t} \) denoting the number of fields.

The fourth objective is the size of the solutions. Smaller solutions can be checked for physical consistency more easily and are computationally less expensive. Incorporating the size as objective further provides information on the dependency between the quality and the complexity of the solutions.

3.3. Pareto optimality

Pareto optimality is a concept to define optimality when dealing with multiple objectives. The state of an economic system is called Pareto optimal when economic resources are distributed such that it is impossible to improve the situation of one person without deteriorating the situation of at least one other person.

For optimization problems that involve multiple, potentially conflicting objectives, usually there exists a set of alternative solutions in which no solution is optimal in the sense that it is superior to all other solutions when considering all objectives. The solutions which are optimal in the sense that there is no other solution which is better concerning all objectives, are called Pareto optimal.

The multiple objectives correspond to different quality criteria of the desired solution. We denote the objective space containing all objective functions as \( \epsilon \) and the solution space containing all potential solutions as \( \mathcal{C} \). A solution \( s_{j} \in \mathcal{C} \) is calculated by comparing prediction (downscaled field) \( y^{q} \) and reference (high-resolution model output) \( y^{s} \). The prediction results from applying the candidate solution (downscaling rule) \( x \in \epsilon \) to the vector of predictors \( x \). Thus, incorporating all dependencies, we can write \( s_{j}(y^{q}, y^{s}) = s_{j}(x, y^{q}, y^{s}) \). For simplicity we denote \( s_{j}(x, y^{q}, y^{s}) = s_{j}(x) \).

Let \( s(x) = (s_{1}(x), s_{2}(x), ..., s_{m}(x)) \) be the objective vector (i.e., the vector containing all \( m \) objectives). The multi-objective minimization problem can then be written as,

\[
s(x) = (s_{1}(x), s_{2}(x), ..., s_{m}(x))^T = \min
\]

(7)

Let us consider two solutions \( a, b \in \epsilon \). The solution \( a \) is said to dominate \( b \) (i.e., \( a > b \)) if

\[
\forall i \in \{1, 2, ..., m\} : s_{i}(a) \leq s_{i}(b) \land \exists j \in \{1, 2, ..., m\} : s_{j}(a) < s_{j}(b).
\]

(8)

In other words, \( a \) dominates \( b \) if \( a \) is at least as good as \( b \) with respect to all objectives, and there exists at least one objective for which \( a \) is better than \( b \).

The solution \( a \) is said to cover \( \hat{b} \) (i.e., \( a \geq \hat{b} \)) if \( a > \hat{b} \) or \( s(x) = s(\hat{b}) \) (i.e., either \( a \) dominates \( \hat{b} \) or they both perform equally well concerning all objectives). The solutions that are not dominated by any of the elements in the solution space \( \epsilon \) are called Pareto optimal.

Fig. 2 shows an example of a minimization problem with two objectives. The squares denote the set of Pareto optimal solutions,
as for each square there is no solution that is better with respect to both objectives (i.e., no solution with smaller $s_1$ and $s_2$). The circles correspond to the non-optimal solutions. For each circle exists at least one solution that is better with respect to both objectives. The numbers will be explained in Section 3.4.1.

3.4. Genetic Programming

Genetic Programming is a machine learning method that “addresses the problem of automatic programming, namely, the problem of how to enable a computer to do useful things without instructing it, step by step on how to do it” (J. Koza in Banzhaf et al., 1997). GP is one of several methods within the area of evolutionary computation. These methods are inspired by the concept of natural evolution. In nature an individual is exposed to environmental pressure. Its chance to survive, reproduce and consequently contribute to the next generation is dependent on its fitness with respect to the environmental conditions. Within the GP framework, a solution evolves in a similar manner as species evolve in nature, i.e., a solution is developed over several generations, each consisting of a large number of candidate solutions.

The candidate solutions (e.g., potential downscaling rules), also called individuals in analogy to the evolution terminology, are composed of program code. In classical (tree-based) Genetic Programming the individuals are represented by parse trees. Fig. 3 shows a simple example of a parse tree consisting of 7 nodes arranged on 4 levels. The parse tree in Fig. 3 embodies an equation consisting of arithmetic functions ($+,−,×,/)$, variables ($a,b,c$) and one constant (1). A parse tree is read starting from the bottom, i.e., in Fig. 3 (b) and (1) serve as input arguments to the subtraction, the output $(b−1)$ and (a) serve as input arguments to the multiplication and so on. The size of the solution is defined as the number of nodes in the parse tree, i.e., the parse tree in Fig. 3 has the size 7.

In general a parse tree consists of functions and terminals. The set of functions and terminals used in a GP run is typically defined by the user and adapted to the problem to be solved. A function set can contain for instance arithmetic functions ($+,−,×,/)$, transcendental functions ($\log,\sin,\ldots$), or conditional statements (if then else, ...). Terminals do not have input arguments and thus terminate the branches of the tree. Terminals serve as fundamental input to the functions. The terminal set can include variables, constants and zero-argument functions, such as a random number generator.

The candidate solutions forming the initial generation of a GP run are automatically and often randomly generated and then tested on the given problem. Each following generation evolves by applying so-called genetic operators, which recombine and modify individuals from the preceding (parent) generation. The better a candidate solution solves a given problem, the greater the chance to contribute to the new generation (e.g., Banzhaf et al., 1997; Eiben and Smith, 2003).

To run GP, one, generally speaking, needs a set of training data, function and terminal sets to build the candidate solutions, and fitness measures (objectives) to quantify the quality of the candidate solutions. Additionally, the user can specify control parameters, such as the population size. Finally, a stopping criterion that ends the evolution must be provided, for instance a maximum number of generations.

3.5. Multi-objective Genetic Programming

The multi-objective fitness assignment according to SPEA requires two main changes compared to traditional GP. Firstly, each generation is split into two sets, $P$ and $P'$. The population $P$ is evolving over time as in traditional GP, whereas the second population $P'$, the so-called Pareto set, contains all Pareto optimal solutions. Secondly, the fitness calculation for individuals in both populations $P$ and $P'$ is based on a comparison between the individuals (i.e., based on the number of individuals dominated by or dominating a solution) rather than on the absolute performance. The fitness assignment is explained in the following.

3.5.1. Fitness assignment

In SPEA the fitness assignment consists of two steps.

(1) To each solution in the Pareto set, $a \in P$, a real value called fitness $f'(a) \in (0,1)$ is assigned. The fitness $f'(a)$ is proportional to the number of individuals $\beta \in P'$ that are covered by $a$, i.e., $a \geq \beta$. Let $N$ be the total number of individuals in $P$. Then $f'(a)$ is defined as

$$f'(a) = \frac{n(\beta | a \geq \beta, a \in P', \beta \in P)}{N + 1} \quad (9)$$

To clearly separate between the fitness of individuals in $P$ and $P'$, the fitness of the individuals in $P'$ is also called strength, hence the name Strength Pareto Evolutionary Algorithm or Short Strength Pareto Approach. The fitness $f(\beta)$ of an individual in the population $P, \beta \in P$, is calculated as the sum over the fitness of all individuals in the Pareto set, $a \in P$, that cover $\beta$,

$$f(\beta) = 1 + \sum_{\alpha, \alpha \geq \beta} f'(a) \quad (10)$$

where $f(\beta) \in [1,N]$. One is added to the sum to ensure that the individuals in the Pareto set $P'$ have better fitness than those in $P$.

Fig. 2 illustrates one possible scenario of a minimization problem with only two objectives $s_1$ and $s_2$. The values indicate the fitness. The circles correspond to the individuals in $P$, the squares to the individuals in $P'$. The lowest point in Fig. 2 shows an individual contained in the Pareto set $P'$ that dominates 3 out of the 7 individuals in $P$. Therefore its fitness $f'$ equals $3/(7+1) = 3/8$. The next lowest point represents an individual from the population $P$, which is dominated only by one individual with a fitness $f'$ of $3/8$. Hence its fitness $f$ calculates as $1+3/8 = 11/8$.

3.5.2. Algorithm

Fig. 4 shows a basic flowchart of the GP algorithm incorporating fitness assignment according to SPEA, which we explain now step by step.
(1) An initial population $\mathcal{P}$ of candidate solutions (individuals) is generated. The initial population can be created randomly or include known approximate solutions of the given problem.

(2) Each candidate solution (i.e., potential downscaling rule) is applied to the training data set.

(3) From the result of (2) the objectives are calculated.

(4) The Pareto set $\mathcal{P}$ is updated: All individuals in population $\mathcal{P}$ that are not dominated within $\mathcal{P}$ are moved to the Pareto set $\mathcal{P}$. The individuals in $\mathcal{P}$ that are covered by another member of $\mathcal{P}$ are removed. In case the number of individuals stored in $\mathcal{P}$ exceeds the allowed maximum, the number of individuals in $\mathcal{P}$ is reduced by hierarchical clustering. In hierarchical clustering (e.g., Hastie et al., 2009) a set of individuals is grouped into clusters according to their similarity. In agglomerative clustering, one variant of hierarchical clustering, initially each individual represents its own cluster. The individuals or later clusters, which are closest according to some distance metric (we use the Euclidian distance in the objective space), are stepwise grouped together to form new, larger clusters. The clustering procedure is stopped when the desired number of clusters, i.e., the maximum size of the Pareto set, is reached. From each cluster, the member closest to the center of the respective cluster is to become part of the new, pruned Pareto set. To assure that all objectives are considered equally, we scale the values before clustering. Since our objectives $s_i$ decrease with better fit, the objectives are scaled via $s^b_i(\alpha) = (s_i(\alpha) - \min_{\beta} s_i(\beta)) / \max_{\beta} s_i(\beta)$, i.e., from the objective we subtract the minimum and divide the result by the maximum occurring in the current Pareto set. Since the fitness results from a greater than/smaller than comparison, scaling does not effect the fitness. Note that scaling is only applied for the clustering. The clustering helps to preserve the diversity of the solutions while shrinking the Pareto set.

(5) The fitness of each individual in $\mathcal{P}$ and $\mathcal{P}^*$ is calculated by comparing the individual’s performances (For details see Section 3.5.1).

(6) If the stopping criterion is met, the final Pareto set is returned. If the stopping criterion is not met, the algorithm continues with (7).

(7) The next generation is created by combining and mutating individuals from the current $\mathcal{P} + \mathcal{P}^*$. The creation of the new generation consists of two steps. First, a sampling procedure is applied to determine the parents. Second, genetic operators (crossover, mutation) are applied to create new individuals.

(7.1) For sampling we use the lexicographic parsimony pressure (Luke et al., 2002) as it is implemented in GPLAB. A number of individuals is randomly drawn from the current $\mathcal{P} + \mathcal{P}^*$. The individual drawn with the best fitness is to become parent. In case several individuals are equally fit the smallest one, i.e., the one consisting of least nodes, is chosen.

(7.2) The genetic operators are applied as follows: Crossover recombines two parents. The parent parse trees are cut at randomly chosen nodes and the separated subtrees are exchanged. (Subtree-)mutation cuts a randomly chosen subtree from the parent and replaces it by a new, randomly created subtree. Parent selection and application of genetic operators are repeated until the new generation is full, i.e., until the population size defined in the settings is reached.

Starting from (2) the succeeding steps are iteratively repeated until the stopping criterion is met (see (6)).

4. Downscaling near-surface temperature

We apply multi-objective GP to the downscaling of near-surface temperature at 10 m height, which is the center of the lowest atmospheric layer in the COSMO model. In a well mixed ABL, temperature anomalies are in good approximation proportional to topographic height (altitude) anomalies (e.g., Schomburg et al., 2010). Under clear sky conditions during night time cold air flows from higher altitudes into the valleys (e.g., Barr and Orgill, 1989) creating pronounced channel structures in the temperature field. Especially the anomalies caused by cold air drainage can grow very large compared to the anomalies in a well-mixed boundary layer (Fig. 5). Therefore, it is important to also capture such more complex processes that cannot be modeled by a linear regression with the downscaling rules.

4.1. Setup

We use a cross-validation approach (i.e., leave-one-out) to test
for overfitting. As mentioned in Section 2 the complete data set contains 8 simulation periods (see Table 1). For computational reasons we extract four time steps (two during the night, two during the day) from each simulation period for training and validating the downscaling rules. By always omitting one simulation period in the training step, the cross-validation approach leaves us with 8 GP runs in total.

The selection of potential predictors included in the terminal is based on our understanding of atmospheric processes, which influence near-surface temperature: Firstly, the coarse fields of near-surface temperature, vertical temperature gradients of the lowest 25 m, 60 m, and 110 m (measures for atmospheric stability), horizontal wind speed at 10 m, vertical wind speed at 20 m, and net radiation at the land surface; secondly fine-scale information on topographic height, plant cover, surface roughness length and surface albedo, as well as a few parameters derived from the spatial structure of the topography field. The latter provide information on local topography relative to its direct surroundings. The predictors are listed in Table 2.

The essential GP settings are summarized in Table 3. We run 200 generations with 100 individuals each, i.e., each run evaluates 20000 potential downscaling rules. The maximum Pareto set size is set to 50. For computational reasons and to keep the solutions readable, we furthermore limit the tree size to 5 levels. Besides the predictors described above, the terminal set contains also random numbers drawn from the interval [0,1]. The function set contains the arithmetic functions with two input arguments and an if statement with four input arguments (i.e., if $a > b$ do $c$ else do $d$).

4.2. Results

Fig. 6 shows the relative reduction of the different objectives for the Pareto optimal downscaling rules. Only the first three objectives, RMSE, ME(STD) and IQD, are included in Fig. 6. All objectives are formulated as penalties, i.e., the smaller the objective the better. The relative reduction of a downscaling rule $a$ with respect to an objective $s$ is given as $\delta_s(a) = 1 - s(a)/s(0)$, where $s(0)$ is the objective when predicting zero anomalies, which corresponds to the spline-interpolated field. The definition of the relative reduction is analogue to a skill score, with the objective for a perfect solution equal to zero. A positive $\delta_s(a)$ indicates that the downscaled field is better than the spline-interpolated field concerning objective $s$; for a perfect downscaling $\delta_s(a) = 1$; for a downscaling that is as good as the interpolated field $\delta_s(a) = 0$; for a downscaling worse than the interpolated field $\delta_s(a) < 0$. Note that $\delta_s(a) \in (-\infty, 1]$, i.e., in the positive direction $\delta_s(a)$ can not exceed 1, whereas it can grow very large in the negative direction.

Fig. 6 indicates very different improvements for the three objectives. Both IQD and ME(STD) are improved by the GP solutions compared to the spline-interpolation (i.e., positive relative reduction). The relative reduction of the IQD amounts on average to 50 – 60% with a maximum of about 90%. The relative reduction of the ME(STD) is slightly lower with an average of about 40 – 50% with a maximum of about 70%. In contrast, the RMSE is on average increased by about 10% compared to the spline-interpolation. At first glance it is dissapointing that the RMSE is not decreased by most of the downscaling rules. Also with the nonlinear regression a pixel wise reproduction of the high-resolution fields appears to be impossible. Nevertheless, it is possible to significantly improve the reproduction of the spatial variability on the subgrid-scale as ME(STD) and IQD are clearly reduced by almost all of the downscaling rules.

In most subfigures of Fig. 6 some outliers are visible. The outliers are most apparent for the RMSE. Most outliers already show a less satisfying performance on the training data sets and tend to stick out in the scatter plots (e.g., Fig. 6a). These solutions are very small (sometimes consisting of only one node). These solutions are part of the Pareto set due to their good performance concerning the 4th objective, i.e., the solution size.

Fig. 7 complements Fig. 6. Shown are 2D-scatter plots for all objectives, now incorporating also the solution size (bottom row), for the downscaling rules validated on October 14th 2007. A minimization of ME(STD) and IQD is not compatible with a minimization of the RMSE (Fig. 7a,b). IQD and ME(STD) are minimized by similar rules (Fig. 7c), which is reasonable as both are aimed at a reproduction of the spatial variability. The performance with respect to the RMSE and the solution SIZE appear to be almost uncorrelated (Fig. 7d). For IQD and ME(STD) there is some

---

### Table 2

<table>
<thead>
<tr>
<th>Atmospheric information (coarse)</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T$</td>
<td>Near-surface temperature</td>
</tr>
<tr>
<td>$T_{25}$</td>
<td>Vert. temp. gradient of lowest 2 layers ($= 25$ m)</td>
</tr>
<tr>
<td>$T_{60}$</td>
<td>Vert. temp. gradient of lowest 3 layers ($= 60$ m)</td>
</tr>
<tr>
<td>$T_{110}$</td>
<td>Vert. temp. gradient of lowest 4 layers ($= 110$ m)</td>
</tr>
<tr>
<td>$W_2$</td>
<td>Near-surface horizontal windspeed</td>
</tr>
<tr>
<td>$W_4$</td>
<td>Near-surface vertical windspeed</td>
</tr>
<tr>
<td>$R_{Lat}$</td>
<td>Net radiation</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Surface information (high-resolution)</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$hsurf$</td>
<td>Topographic height</td>
</tr>
<tr>
<td>$hsurf_a$</td>
<td>Topographic height anomaly</td>
</tr>
<tr>
<td>$topo$</td>
<td>Mean height difference to neighboring grid points</td>
</tr>
<tr>
<td>$topo_a$</td>
<td>Anomaly of $topo$</td>
</tr>
<tr>
<td>$topo_1$</td>
<td>Slope to lowest neighboring grid point</td>
</tr>
<tr>
<td>$topo_2$</td>
<td>Slope to highest neighboring grid point</td>
</tr>
<tr>
<td>$topo_4$</td>
<td>Number of direct neighbors lower than grid point</td>
</tr>
<tr>
<td>$plc$</td>
<td>Plant cover</td>
</tr>
<tr>
<td>$z_0$</td>
<td>Roughness length</td>
</tr>
<tr>
<td>$alb$</td>
<td>Albedo</td>
</tr>
</tbody>
</table>

---

### Table 3

Summary of the GP settings. Protected division means that division by zero returns the dividend not an error.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Function set</td>
<td>$+,-,*,$ protected; if</td>
</tr>
<tr>
<td>Terminal set</td>
<td>Random numbers [0,1], variables (Table 2)</td>
</tr>
<tr>
<td>Generations</td>
<td>200</td>
</tr>
<tr>
<td>Population size</td>
<td>100</td>
</tr>
<tr>
<td>Max. Pareto set size</td>
<td>50</td>
</tr>
<tr>
<td>Genetic operators</td>
<td>(subtree-*)mutation, crossover</td>
</tr>
<tr>
<td>Max. tree levels</td>
<td>5</td>
</tr>
</tbody>
</table>

---

**Fig. 5.** Diurnal cycle of the root mean square error (RMSE) of the spline-interpolated fields w.r.t. the high-resolution reference fields. May 10th 2008 was a cloudy day. August 28th 2008 was a cloudy day with some rain.

**Fig. 6.** shows the relative reduction of the different objectives for the Pareto optimal downscaling rules. Only the first three objectives, RMSE, ME(STD) and IQD, are included in Fig. 6. All objectives are formulated as penalties, i.e., the smaller the objective the better. The relative reduction of a downscaling rule $a$ with respect to an objective $s$ is given as $\delta_s(a) = 1 - s(a)/s(0)$, where $s(0)$ is the objective when predicting zero anomalies, which corresponds to the spline-interpolated field. The definition of the relative reduction is analogue to a skill score, with the objective for a perfect solution equal to zero. A positive $\delta_s(a)$ indicates that the downscaled field is better than the spline-interpolated field concerning objective $s$; for a perfect downscaling $\delta_s(a) = 1$; for a downscaling that is as good as the interpolated field $\delta_s(a) = 0$; for a downscaling worse than the interpolated field $\delta_s(a) < 0$. Note that $\delta_s(a) \in (-\infty, 1]$, i.e., in the positive direction $\delta_s(a)$ can not exceed 1, whereas it can grow very large in the negative direction.

**Fig. 7.** complements Fig. 6. Shown are 2D-scatter plots for all objectives, now incorporating also the solution size (bottom row), for the downscaling rules validated on October 14th 2007. A minimization of ME(STD) and IQD is not compatible with a minimization of the RMSE (Fig. 7a,b). IQD and ME(STD) are minimized by similar rules (Fig. 7c), which is reasonable as both are aimed at a reproduction of the spatial variability. The performance with respect to the RMSE and the solution SIZE appear to be almost uncorrelated (Fig. 7d). For IQD and ME(STD) there is some
Fig. 6. Relative reduction of RMSE, ME(STD) and IQD for the 8 Pareto sets. The larger the value for the relative reduction, the better the performance concerning the respective objective. The 8 subfigures show the results of 8 GP runs each omitting a different day in the training. The blue circles indicate the performance of each of the 50 Pareto optimal solutions on the training data and the red crosses indicate the performance on the validation data (i.e., the day omitted in the training); (a) shows the results of the GP run, where August 27th 2007 was omitted for training and so on.
dependency on the solution size. The more complex a downscaling rule from the Pareto set is, the more likely it exhibits a good performance concerning IQD and ME(STD). There is, however, no simple linear relation between solution quality and complexity. There are downscaling rules of intermediate complexity with about 20 nodes, showing a similar performance as the very large solutions with about 30 nodes. Still a certain amount of complexity has to be accepted in order to achieve a good performance concerning ME(STD) and IQD. The subjective rating of the downscaling rules indicated by the colors in Fig. 7 is explained in the second paragraph of the discussion.

Fig. 8 shows the difference in relative reduction of the different objectives between training and validation data set \( (\text{rmse} - s_{\text{val}}) \) for all 8 runs. Each box results from 50 values, one for each solution from the Pareto set. The horizontal line within the boxes is the median, the upper and lower boundaries of the boxes correspond to the 75%- and 25%-quantiles. The whiskers indicate the range spanned by maximum and minimum. The length of the whiskers is restricted to 1.5 \times box size. Values outside this range are considered outliers and shown as circles.

We examine now in more detail the results of one downscaling rule for October 14th 2007, a clear sky day. Figs. 9 and 10 show the performance of one GP downscaling rule at 12:00 UTC and 24:00 UTC, respectively. The rule shown (rule 1) is:

\[
T^* = 0.74 \times h_{\text{surf}a} \times T_{\text{gr110}}
\]

with \( h_{\text{surf}a} \) being the topographic height anomaly, \( T_{\text{gr110}} \) the temperature gradient of the lowest 110 m, i.e., 4 lowest model layers (see Table 2).

The daytime temperature field in Fig. 9 is well predicted by both the GP downscaling rule and the linear regression of Schomburg et al. (2010). There are no significant differences. For the nighttime field shown in Fig. 10 the improvement by GP is striking. The pattern formed by nightly cold air drainage is captured by the GP downscaling rule, while the linear regression approach from Schomburg et al. (2010) fails to predict any temperature anomalies.

Figs. 9 and 10 show that already the very simple GP downscaling rule already seems to provide a good match of the reference fields. However, the RMSE in increased by 4% for the training data and by 25% for the validation data compared to the spline-interpolated field (Table 4). Concerning ME(STD) and IQD we achieve improvements of 40%–86%. For ME(STD) and IQD the rule performs better on the validation day than on the training data set.
Fig. 11 shows the predicted temperature anomalies for the nighttime field on October 14th 2007. Large anomalies are predicted in areas where also the variability in the reference field is strong. However, the RMSE is not reduced, which is obvious from the difference plot (middle figure of second row in Fig. 11). The downscaled anomalies appear much smoother than in the reference field, where local minima and maxima are more pronounced and much sharper.

Fig. 12 illustrates the standard deviation, which ME(STD) aims to optimize. As already shown in Table 4 the ME(STD) is clearly reduced. This improvement is most obvious in regions with distinct topography, i.e., the Eifel region in the south of the domain and the Bergisches Land in the northeastern corner of the domain. Here the variability is both higher and more strongly connected to the topography than in flatter areas. Hence, it is easier to achieve improvements in this regions. The difference map, $\text{std}(T_{\text{ref}}) - \text{std}(T_{\text{downsc}})$, shows that in some grid boxes the variability is slightly overestimated by the downscaling rule. As discussed earlier this might be improved by adding more clear sky situations to the training data set. The underestimation of

<table>
<thead>
<tr>
<th></th>
<th>RMSE</th>
<th>ME(STD)</th>
<th>IQD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>−0.04</td>
<td>0.41</td>
<td>0.64</td>
</tr>
<tr>
<td>Validation</td>
<td>−0.25</td>
<td>0.50</td>
<td>0.87</td>
</tr>
</tbody>
</table>

Table 4: Relative reduction of the different objectives for an example downscaling rule (rule 1); compare Figs. 9 and 10.

Fig. 9 Performance of an example solution (rule 1) from the run omitting October 14th 2007 in the training. Shown is the temperature field at 12:00 UTC on October 14th 2007. The upper figures show the full fields (i.e., 112 × 112 km); the bottom figures show a zoom of an area of 28 × 28 km: (a) shows the spline-interpolated field; (b) shows the field resulting from the downscaling rule by Schomburg et al. (2010); (c) shows the field resulting from the rule found by GP; (d) shows the reference field from the high-resolution model run.

Fig. 10. Same as in Fig. 9, but for October 14th 2007, 24:00 UTC. (a) and (b) are the same, because the linear regression based rule by Schomburg et al. (2010) does not predict any anomalies in stable atmospheres.
variability in regions with less distinct topography might be reduced by replacing \( \text{ME}(\text{STD}) \) by a ratio of standard deviations to assign more weight to grid boxes with less absolute variability.

Fig. 13 shows the difference between the discrete PDFs of the absolute values (not the anomalies) of the full fields in order to illustrate the performance with respect to the IQD. The difference is clearly reduced by the downscaling. However, especially for the lower temperatures around 276–277 K the downscaling has only a small effect. This is again attributed to the very sharp local minima, which the downscaling cannot reproduce appropriately.

So far we only looked at the results for one very simple GP downscaling rule. Actually many of the GP downscaling rules contain the term \( \frac{\text{hsurfa}}{\text{C2}} \times T_{\text{gr}} \) with one of the three temperature gradients offered as potential predictors. This is physically intuitive and reasonable. An example for a more complex downscaling rule is:

\[
T^* = \begin{cases} 
\text{if } \text{hsurfa} > 0.98 \\
\text{then } \text{hsurfa} \times T_{\text{gr}60} \\
\text{else } \text{hsurfa} \times T_{\text{gr}110} \\
\text{if } T_{\text{gr}25} > T_{\text{gr}110} \\
\text{then } W_v \\
\text{else } \text{topo2} \\
\text{else } W_v 
\end{cases}
\]

In the first part, again the surface height anomaly \( \text{hsurfa} \) is multiplied by the temperature gradients \( T_{\text{gr}} \). The second part either adds vertical wind speed \( W_v \) or one of the topography based parameters, \( \text{topo2} \), depending on atmospheric stability. A near-surface temperature gradient \( >0.83 \text{ km}^{-1} \) is very rare, so in nearly all cases \( W_v \) is added. This does not have any effect, because \( W_v \) is constant within each coarse pixel and we keep the mean temperature over the coarse pixels constant, in order not to affect the energy balance by the downscaling. Of course this restriction could be loosened depending on the application.

The solutions consisting of only one node perform comparably bad as one would expect. Predictors selected by GP for these solutions include roughness length, plant cover, and some of the parameters derived from topography as all these parameters capture some part of the surface heterogeneity. None of these can compete with the slightly more complex rules with \( \geq 3 \) nodes.

5. Discussion

As we extensively compare our approach to the downscaling developed by Schomburg et al. (2010, 2012), we briefly discuss the adequacy of this comparison. Making this comparison seems natural as the current study can be seen as an extension to the approach by Schomburg et al. (2010, 2012). We exploit the same data set, firstly interpolate the coarse field and secondly predict the high-resolution anomalies. However, the list of potential predictors
for both algorithms is not identical, thus we cannot directly ascribe the improvements provided by the GP based algorithm purely to the multi-objective approach and its potential nonlinearity. Predictor selection might be just as important. For future comparison one may use more advanced linear models, multivariate linear regression or generalized linear models, as well as other nonlinear methods, for instance generalized additive models like in Malone et al. (2012) or an artificial neural net.

In Section 4.2 we notice that especially in Fig. 10 the downscaled field makes a better visual impression than the values of the objectives in Table 4 suggest, especially concerning the RMSE. To find out what makes the spatial fields ‘look’ similar to the reference, we have categorized all 50 downsampling rules from one Pareto set by visual inspection. The colors of the points in Fig. 7 indicate the results of this subjective evaluation. Good, average and bad solutions can obviously not be clustered easily, but some systematics are obvious. Fig. 7 (d) indicates that a certain level of RMSE must be accepted in order to reasonably reproduce the systematics are obvious. Fig. 7 (d) indicates that a certain level of RMSE must be accepted in order to reasonably reproduce the fine-scale variability, because subjectively good solutions tend to perform worse concerning the RMSE than visually bad solutions. The better a solution performs concerning the ME(STD), the better it seems visually. Concerning the IQD the behavior is similar. When subjectively rating the solutions, more complex (larger) solutions tend to perform better. Thus, a minimum solution size is required to account for the complexity of the processes involved in building the fine-scale structures. However, the visually best looking solution seems not to be detectable from the objectives alone. Fig. 14 shows the downscaled fields of another downsampling rule for the same day and time as in Figs. 9 and 10. The solution in Fig. 14 we would visually rate at most as average, because the spatial structure of the fine-scale variability is not well reproduced. This we could not conclude from the objectives alone, which give quite similar results for both downsampling rules. This suggests to reconsider the objectives to find more appropriate ways to quantify the similarity between the fine-scale patterns of downscaled and reference fields.

We have tested the errors of temporal and spatial correlation as additional objectives, which caused the size of the solutions to increase considerably. The correlation between consecutive time-steps is overestimated by the GP method due to two reasons. First, the coarse or spline-interpolated fields are in general correlated more strongly in time than the high-resolution reference fields. Second, the fine-scale structures are predicted from (on the considered time scale) invariant high-resolution surface properties. The latter is also the reason why the spatial correlation is hard to predict. The reproduction of spatio-temporal correlations might be improved by adding a function creating noise (e.g., adding a simple random number generator to the function set, with an input argument defining the distribution to draw from) or by adding spatio-temporally correlated noise to the downscaled fields. It might be also reasonable to replace the mean error of the standard deviation by a ratio between predicted and reference subpixel standard deviation. In doing so, reproducing the subgrid variability in pixels with less absolute variability (than in mountainous areas for instance) would gain more weight in the fitting procedure.

A temperature downscaling as presented in this study does not necessarily improve the estimation of the surface fluxes, since the objectives used, neighborhood based RMSE, ME(STD) and IQD, do not account for covariances between atmospheric and surface temperature and other atmospheric and surface state variables. Especially for temperature this can be problematic as the turbulent exchange coefficients are nonlinear functions of the near-surface atmospheric stability, which is determined by the difference between the atmospheric temperature in the lowest model layer and the temperature at the surface.

To obtain a preliminary understanding of the impact of the temperature downscaling on the surface fluxes, we have estimated the exchange fluxes of latent heat $\mathcal{L}H$ and sensible heat $\mathcal{S}H$ at the surface $s$ for one clear sky day (October 14th, 2007) as follows:

$$SH_s = - \rho \kappa_L \mathcal{L}H_s (T_s - T)$$  \hfill (11)

$$\mathcal{L}H_s = - \rho \kappa_L \mathcal{L}H_s (Q_s - Q_c^s)$$  \hfill (12)

with the transfer coefficients $\kappa_L$ determined based on Louis (1979) using Monin-Obukhov theory (Monin and Obukhov, 1954), $c$ being the specific heat, $L$ being the latent heat of vaporization, $\rho$ being the air density, $W_h$ being the horizontal wind speed, $T$ and $T_i$ being the temperature and $Q_s$ and $Q_c^s$ the specific humidity in the lowest atmospheric layer and at the surface. All variables required are taken from the high-resolution COSMO runs. To access the effect of the temperature downscaling the near-surface temperature fields are replaced successively by the corresponding spline-interpolated fields and the downscaled fields resulting from either a GP based downsampling rule (rule 1) or the linear regression based rule from Schomburg et al. (2010).

Fig. 15 shows that for October 14th, 2007, a clear sky day, the GP based downsampling rule reduces the RMSE of the fluxes (compared to the spline interpolation) similar to the linear regression. For nighttime situations, however, the error increases slightly for the latent heat flux (5–10% in average over the whole night) and more distinct for the sensible heat flux (about 30% when comparing the fluxes on the coarse scale; about 10% when comparing on the fine scale). One reason for this effect is that the fine scale fields of 10 m-temperature and ground surface temperature can be shifted against each other as the GP based downsampling rules do not reproduce the 10 m-temperature fields grid point wise.

However, we believe that this results can not be transferred to the flux estimation in couple modeling runs or even when driving land-surface models with the downscaled temperature field. In reality and in coupled simulations the ground surface temperature reacts to the atmospheric near-surface temperature. A large temperature difference leads to strong fluxes reducing the difference. And in deed, we see that in our flux estimation the fluxes are by tendency overestimated in their absolute value. When driving a land-surface model such that the ground surface temperature

![Fig. 14. Performance of another example solution validated on October 14th 2007 (same case as shown in Figs. 9 and 10); (a) shows the downscaled field at 12:00 UTC (compare Fig. 9); (b) shows the downscaled field at 24:00 UTC (compare Fig. 10).](image-url)
reacts to the downscaled atmospheric forcing we expect this effect, the overestimation tendencies, to be reduced. On the contrary, even the opposite, an underestimation by tendency, may occur, as the downscaled temperature fields are too strongly correlated in time and thus the ground surface temperature has more time to adjust to the 10 m-temperature as in the high-resolution reference simulations. If we find such tendencies, a better representation of the temporal correlation will be crucial. When explicitly aiming at reproducing the fluxes one could further introduce an objective quantifying near-surface stability. Also other inter-variable relationships can potentially be considered by the objectives. Further, one can add objectives quantifying the relations between atmospheric quantities and surface properties, for instance the correlation between topographic height and temperature anomaly.

6. Conclusion and outlook

We have introduced multi-objective GP for the discovery of rules to downscale mesoscale atmospheric model output (here 2.8 km grid resolution) to the high resolution required for driving land surface and hydrological models (here 400 m). Statistical downscaling of coherent spatial fields is a difficult task compared to time series for a set of local stations. Aiming purely at an optimization of the RMSE, we have not been able to detect reasonable downscaling rules which outperform linear regression. The predicted temperature anomalies were much too small. However, applying multi-objective GP to the downscaling of near-surface temperature, we have shown that this more complex method leads to improvements compared to a conditional linear regression. For instance the pronounced fine-scale variability occurring in regions with distinct topography during nighttime temperature inversions can be reproduced to a great extend with the multi-objective GP algorithm.

An atmospheric downscaling as described in this study can pursue different goals. Ideally it leads to an improved representation of mass and energy fluxes in coupled subsurface-land-surface-atmosphere simulations, but it can also enhance the representation of threshold dependent processes, such as runoff and snow melt, within fully-coupled simulations as well as in stand-alone land surface/subsurface simulations. Further, the downscaled near-surface fields might offer valuable input for agricultural models, for which for instance a good representation of night frost is important. As discussed above the effect of the presented downscaling approach on the simulation of mass and energy fluxes has to be assessed carefully. The detailed analysis of the effect on mass and energy fluxes is beyond the scope of this study and will be a central aspect of future work.

Ongoing work includes an update of the training data set using the most recent version of the COSMO model, and covering more simulation days, especially more clear-sky days and extreme weather conditions. The multi-objective GP is applied to the remaining variables required as atmospheric input to land surface and subsurface models, i.e., precipitation, near-surface specific humidity, near-surface wind speed, long-wave and short-wave radiation. The discovered downscaling rules will be implemented in the integrated modeling platform TerrSysMP by Shrestha et al. (2014), which is composed of the atmospheric model COSMO (Baldauf et al., 2011; Doms and Schättler, 2002), the land surface model CLM (Oleson et al., 2004) and the hydrological model ParFlow (Kollet and Maxwell, 2006). The three component models are coupled via the external coupler OASIS (Valcke, 2013), which currently incorporates the 3-step-downscaling scheme by Schomburg et al. (2010). It is to be seen how a fully coupled modeling system reacts to the atmospheric downscaling. This can be investigated stepwise, starting with comparing land surface and hydrological simulations with the different atmospheric forcings at

![Fig. 15. Diurnal cycle (on October 14th, 2007) of the relative reduction of (pixel wise) root mean square error and root mean square error of the subgrid-scale standard deviation for latent (LH) and sensible heat (SH) flux resulting from different downscaling rules for the near-surface temperature. ‘lin reg’ refers to the linear regression based downscaling rule from Schomburg et al. (2010), ‘GP’ (rule 1) refers to the downscaling rule from Figs. 9 and 10. The (pixel wise) RMSE has been calculated on two scales: for the high-resolution fields (high-res.) and after averaging the fluxes to the coarse scale (coarse).](image-url)
coarse scale, downscaled and from high-resolution atmospheric model runs. Potential feedbacks can be studied in a second step with the fully coupled TerrSysMP.

Finally, multi-objective GP generates a set of Pareto optimal solutions and not just one downscaling rule, which suggests the implementation of a downsampling ensemble by either using different rules in each ensemble run or by randomly switching between rules taking into account temporal consistency. An ensemble approach could also help to estimate the sources of uncertainty induced by the downsampling procedure and might ultimately lead to a stochastic parameterization scheme for the near-surface atmospheric subgrid-scale variability.

Acknowledgments

The authors would like to thank the three anonymous reviewers, whose comments and suggestions helped to substantially improve on earlier versions of the manuscript. We further thank A. Schomburg for providing the high-resolution COSMO model output which is is available at http://www.tr32db.uni-koeln.de. This work has been carried out in the framework of the Transregional Collaborative Research Centre 32 on “Patterns in Soil-Vegetation-Atmosphere-Systems” funded by the Deutsche Forschungsgemeinschaft (DFG).

References


Baldauf, M., Seifert, A., Fritsche, T., Zieren, T., 2015. Potential feedbacks can be studied in a second step with the fully coupled TerrSysMP.


Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, vol. 2. Springer,


