Genetic Programming for Estimation of Heat Flux between the Atmosphere and Sea Ice in Polar Regions

Karolina Stanislawska
Finnish Meteorological Institute
00560 Helsinki, Finland
karolina.stanislawska@fmi.fi

Krzysztof Krawiec
Poznan Univ. of Technology
60965 Poznań, Poland
krawiec@cs.put.poznan.pl

Timo Vihma
Finnish Meteorological Institute
00560 Helsinki, Finland
timo.vihma@fmi.fi

ABSTRACT

The Earth surface and atmosphere exchange heat via turbulent fluxes. An accurate description of the heat exchange is essential in modelling the weather and climate. In these models the heat fluxes are described applying the Monin-Obukhov similarity theory, where the flux depends on the air-surface temperature difference and wind speed. The theory makes idealized assumptions and the resulting estimates often have large errors. This is the case particularly in conditions when the air is warmer than the Earth surface, i.e., the atmospheric boundary layer is stably stratified, and turbulence is therefore weak. This is a common situation over snow and ice in the Arctic and Antarctic. In this paper, we present alternative models for heat flux estimation evolved by means of genetic programming (GP). To this aim, we utilize the best heat flux data collected in the Arctic and Antarctic sea ice zones. We obtain GP models that are more accurate, robust, and conceptually novel from the viewpoint of meteorology. Contrary to the Monin-Obukhov theory, the GP equations are not solely based on the air-surface temperature difference and wind speed, but include also radiative fluxes that improve the performance of the method. These results open the door to a new class of approaches to heat flux prediction with potential applications in weather and climate models.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning—*Induction*; I.6.3 [**Simulation and Modeling**]: Applications

Keywords

Modeling; genetic programming; symbolic regression; meteorology

GECCO '15, July 11 - 15, 2015, Madrid, Spain

ACM ISBN 978-1-4503-3472-3/15/07…$15.00

DOI: http://dx.doi.org/10.1145/2739480.2754675

1. INTRODUCTION

The net heat flux (NF) at the Earth surface is of critical importance for the energetic balance of our planet and consist of the following components:

$$NF = (1-\alpha)SWR_{\downarrow} + LWR + H + LE + C, \qquad (1)$$

where $SWR_{\downarrow}$ is the downward solar shortwave radiation, and α is the fraction of it that is reflected from the surface (called as the surface albedo). LWR is the difference of the downward thermal longwave radiation from the atmosphere ($LWR_{\downarrow}$) and the upward longwave radiation emitted by the Earth surface ($LWR_{\uparrow}$). H and LE are the turbulent fluxes of *sensible heat* and *latent heat*, respectively. The latent heat flux is the flux of water vapour, due to evaporation or condensation, multiplied by the latent heat of evaporation. The last term, C, is the heat flux to/from the layers below the Earth surface. All fluxes are defined positive towards the surface. $SWR_{\downarrow}$ depends above all on the solar zenith angle and clouds, and α mostly depends on the physical properties of the surface. The downward longwave radiation depends on the temperature and emissivity profiles of the atmosphere, clouds being the most important factor affecting the latter, and the longwave radiation emitted by the Earth surface depends on the surface temperature and emissivity. On solid surface types, C is due to heat conduction. The turbulent fluxes H and LE are driven by the surface-air differences in temperature and specific humidity, respectively.

As a global annual average, the heat budget of the Earth surface is closed so that the the main heat input is $SWR_{\downarrow}$, balanced by heat loss via LWR, H, and LE. C is typically a small term being positive during night/winter and negative during day/summer. During night or winter in polar regions, however, there is no solar radiation, and the main heat input to the Earth surface is via H and C, balanced by heat loss via LWR. The heat fluxes control the evolution of the Earth surface temperature: a positive NF results in warming and a negative one in cooling. The surface temperature is, however, not affected if the positive NF is entirely used to melt ice or snow, or a negative NF causes freezing of water surfaces. One of the most striking signals of the recent climate warming has been the extensive reduction of the Arctic sea ice [3] and Arctic and mid-latitude terrestrial snow cover [5]. The larger heat input from the atmosphere to snow and ice have had a dominating role in the snow and ice melt [6].

To successfully model the evolution of weather and climate, the terms in (1) need to be accurately described.

The problem is that the heat fluxes are generated by small-scale physical processes (radiation, turbulence, molecular heat conduction) but weather and climate models have to cover large geographical areas, and therefore have a coarse resolution. A typical horizontal grid spacing in numerical weather prediction (NWP) models, which form the basis of short-term weather forecasts, is of the order of 10 km and for climate models of the order of 100 km. The vertical grid spacing is smallest in the lowest layers of the atmosphere, so that the lowest atmospheric model level is typically 10-50 m above the Earth surface, which itself forms another level of the model. Hence, the heat fluxes need to be described as functions of the variables that are resolved by the model grid, i.e., via so-called *subgrid-scale parameterization*. The resolved variables include wind speed and direction, air temperature and humidity, as well as cloud water/ice content. Likewise, at the Earth surface, the model calculates the temperature and humidity.

Particularly in polar regions, the parameterization of all terms in (1) includes challenges [17]. In the case of radiative fluxes these are mostly related to cloud physics, and in the case of the conductive heat flux to the snow and ice properties. Handling the turbulent fluxes, H and LE, is most challenging when the air is warmer than the snow/ice surface, i.e. the lowermost atmosphere is statically stably stratified. Then the turbulent fluxes are small and liable to parameterization errors. As there are not many accurate observations available on the latent heat flux, in this paper we focus on the parameterization of sensible heat flux over Arctic and Antarctic sea ice. We review the main problems in conventionally applied parameterization methods (Section 2), discuss the factors that affect H and argue about the suitability of using genetic programing for flux modeling (Section 3), and describe two high-quality data sets that capture the dynamics of the sensible heat flux in the polar regions (Section 4). The original contributions of this study start with Section 5, where we present a novel approach to heat flux modeling based on genetic programming. Section 6 presents and discusses the evolved models, which we then compare to the conventional parameterization in Section 7, to conclude the paper with Section 8.

2. LIMITATIONS OF CONVENTIONAL MODELS OF FLUX PREDICTION

In climate and NWP models, the parameterization of the turbulent flux of sensible heat H is based on the classical Monin-Obukhov theory [11]. According to the theory, H depends on the vertical profile of potential temperature, the momentum flux, and a parameter describing the static stability. This yields a practical equation for H, the so-called *bulk formula*:

$$H = \rho c_p C_H (\theta_s - \theta_a) V, \tag{2}$$

where ρ is the air density, c_p is the specific heat, C_H is the turbulent heat exchange coefficient, θ_s is the surface potential temperature, θ_a is the air potential temperature, and V is the wind speed. The C_H is calculated on the basis of the stability parameter and two parameters characterizing the surface roughness from the point of view of turbulent mixing. The Monin-Obukhov theory includes, however, idealized assumptions: horizontal homogeneity, semi-stationarity, and that the turbulent flux is constant in the vertical from the Earth surface up to the lowest atmospheric model level. In conditions of stable stratification, when turbulence only exists in a shallow layer close to the Earth surface, the last assumption is highly idealized. Further, the effect of the static stability of C_H can only be determined experimentally, and in conditions of stable stratification there is a lot of scatter between various experimental formulae for C_H. Observations indicate that when the stratification is very stable, the magnitude of H may decrease although the air-surface potential temperature difference, $\theta_s - \theta_a$, increases [10]. This is due to the dominating effect of decreasing C_H, but the sensitive interaction is very difficult to reproduce in models.

We note that the results for H based on (2) depend on how the surface roughness effects and stability dependence of C_H are parameterized. In this study, we calculate H according to (2) using the roughness and stability parameterization as in the operational NWP model of the European Centre for Medium-Range Weather Forecasts (ECMWF)[1]. Hereafter, the ECMWF method of calculating H is referred as *conventional model*.

Hence, NWP and climate models typically have their largest errors in surface and near-surface air temperature under stable stratification in high latitudes. Even in short-term forecasts by NWP models the errors of estimation of 2-m air temperature may reach 10K [2]. In experiments involving six regional climate models for the Arctic sea ice zone, the observed and modelled H did not correlate [16]. In general, in conditions of stable stratification, models tend to overestimate H [15]. These problems urgently call for new approaches for parameterization of H.

3. GENETIC PROGRAMMING FOR DETERMINING HEAT FLUX

Given the difficult nature of the turbulent flux of the sensible heat and the unsatisfactory performance of the conventional models, in this paper we resort to genetic programming (GP) as a means for synthesizing a heat flux model from experimental data. In doing so, we anticipate obtaining models that (i) at least tie with the conventional models with respect to error on experimental data, (ii) are to some extent transparent/interpretable, and by this token (iii) possibly cast new light on the nature of the complex phenomenon of heat flux.

Genetic programming is a heuristic, stochastic approach to program synthesis. GP synthesizes a program in a given domain-specific language by maintaining a working population of programs (candidate solutions) and manipulating them by means of search operators. This iterative search process is driven by a domain-specific fitness function, which evaluates the quality of candidate solutions and so imposes certain selection pressure; this in turn causes some programs to be appointed as parents and so give rise to the subsequent generation of models. Given an appropriately designed domain-specific language, GP can be also used to induce functional models of dependencies between multiple observables, i.e., independent input variables that form program input and a dependent output variable that is to be predicted by a program. Applying GP in this mode is also known as *symbolic regression*.

[1]Details of the ECMWF scheme can be found at http://old.ecmwf.int/research/ifsdocs/CY40r1/IFSPart4.pdf (Section 3.2).

In our search for alternative parameterizations of the turbulent flux of sensible heat H, we not only rely on the non-orthodox method of model synthesis (GP), but consider also all potentially relevant variables, including those absent in the traditional parameterization (2). One approach is to regard $SWR_{\downarrow}$ and $LWR_{\downarrow}$ as external drivers for the atmosphere-surface heat exchange and H, LE, C, and $LWR_{\uparrow}$ as fluxes that respond to the external drivers, all depending on the surface temperature. This approach could yield parameterization schemes based on $SWR_{\downarrow}$ and $LWR_{\downarrow}$ (or on the factors controlling these radiative fluxes, such as the solar zenith angle and cloud cover). Another approach would be to keep the potential temperature difference $(\theta_s - \theta_a)$, as in (2), but abandon the highly idealized Monin-Obukhov similarity theory. The dependence of H on $(\theta_s - \theta_a)$ and various other meteorological variables could be tested experimentally. It is also possible to combine the above-mentioned approaches, and parameterize H on the basis of both radiative fluxes and the variables used in (2).

In this paper, we delegate the choice of relevant variables almost entirely to GP, providing only three compound inputs for the models (differences of temperature and radiation). The choice of the variables is to an extent dependent also on the availability and quality of experimental data, which we cover in the next section.

4. EXPERIMENTAL DATA ON HEAT FLUX

To drive the GP evolution we use the best and temporally most extensive meteorological datasets ever collected from the Arctic and Antarctic sea ice zones. These originate from projects Ice Station Weddell (ISW) in 1992 and the Surface Heat Budget of the Arctic Ocean (SHEBA) in 1997-1998. Since then the data from these field campaigns have been widely used by polar scientists, as there have not been other as comprehensive long-term measurement campaigns over polar sea ice since ISW and SHEBA.

ISW was a joint US-Russian project to investigate the atmosphere, sea ice, and ocean in the Weddell Sea, Antarctic [1]. The station was located on a sea-ice flow drifting northwards from 72°S up to 66°S from February to June 1992. The station location deep in the Antarctic sea ice zone and its operation period in austral autumn and early winter gave an excellent opportunity to make observations on the stable boundary layer. In total, over 2,000 hours of nearly continuous data on radiation, temperature, sensible and latent heat, wind speed and other variables were collected.

SHEBA was a year-round campaign to take comprehensive measurements that would help understanding processes that drive the surface energy budget in the Arctic [12]. The experiments spanned from October 1997 to September 1998 and were performed on an ice floe drifting in the Beaufort and Chukchi Seas in the Arctic, between 74°N and 81°N.

The SHEBA and ISW datasets contain respectively 6,004 and 1,024 hourly resolved observations with valid measurements of all variables. To monitor the extent of overfitting in our GP experiments, we split these datasets into training and test parts: approximately 50 percent of randomly selected observations form the former, and the remaining ones the latter. By relying on random partitioning rather than splitting the timeline into two continuous time intervals, we hope to avoid the risk of biasing the training process towards a particular time interval. This preprocessing resulted in the SHEBA dataset being partitioned into 3,004 training observations and 3,000 testing observations, and ISW partitioned into, respectively, 515 and 509 observations. The dataset variables are listed in the top part of Table 1.

Table 1: The terminal nodes (leaves of expression trees) used in the evolved models.

Original inputs (the variables present in the datasets)	
ShortIn	incoming (downward) shortwave radiation
ShortOut	outgoing (upward) shortwave radiation
LongIn	incoming (downward) longwave radiation
LongOut	outgoing (upward) longwave radiation
SurfTemp	surface temperature
AirTemp	air temperature
WindSpeed	wind speed
SpecHum	specific humidity
Derived inputs (compounds of original variables)	
NetShort	incoming minus outgoing shortwave radiat.
NetLong	incoming minus outgoing longwave radiat.
TempDiff	air minus surface temperature
Constants	
CKarman	von Karman constant = 0.405
CSpecHeat	specific heat of air = 1004
ERC	constant drawn uniformly from [-100,100]

Note that although heat flux is naturally a temporal phenomenon, and each observation is associated with a specific timestamp, it is widely agreed in meteorology that H should be parameterized *diagnostically*, i.e., using the simultaneous values of other variables (rather than *prognostically*, i.e., to predict the future values of H). We adhere to that convention in this study. The models we discuss in the subsequent sections are thus time-free in the sense that they express the momentary dependency of the output variable on the input variables. In other words, we pose the problem as a conventional regression problem in terms of statistics and machine learning.

5. EXPERIMENTAL SETUP

We rely on the tree-based GP [7], the genre of GP that is arguably the most natural for handling functional (side effect-free) expressions. A single model is an expression tree that fetches the values of input variables (independent variables) and returns the predicted value of the dependent variable. The terminals (tree leaves), presented in Table 1, include the input variables provided in the SHEBA and ISW datasets (Section 4), simple compounds of them (radiation and temperature differences), and constants (random and domain-specific). The set of instructions (inner nodes of expression trees) comprises arithmetic operators (+, -, *, /), inversion, negation, and selected transcendental functions (*sqr*, *sqrt*, *log*, x^y). Where needed, the functions are protected against invalid arguments. We anticipate that inclusion of the transcendental functions may be essential, given the prevalence of various forms of nonlinearity in physical phenomena governing the turbulent heat flux.

All GP configurations considered in this paper implement generational evolutionary workflow with fairly conventional settings. Parent models are appointed via tournament of size 7. A new model is built from them via subtree-swapping crossover (with probability 0.5), subtree-replacing mutation (0.1), ERC constant mutation (0.3) or cloning (0.1). The

Table 2: Mean absolute error (MAE) of the best-of-run models obtained with particular approaches (for GP, averaged over 30 evolutionary runs and accompanied by 0.95 confidence intervals). All results in W/m^2.

	SHEBA dataset (Arctic)				*ISW dataset (Antarctic)*			
Configuration	*Training MAE*		*Test MAE*		*Training MAE*		*Test MAE*	
Conventional model	3.7568		3.7186		12.6011		11.9325	
LR	4.6465		4.5882		6.0205		6.5886	
GP-MAE	3.3690	[3.286, 3.452]	3.3213	[3.239, 3.404]	5.4819	[5.341, 5.623]	6.0480	[5.837, 6.259]
GP-COR	3.0764	[3.037, 3.115]	3.0213	[2.980, 3.063]	5.2207	[5.179, 5.262]	5.9091	[5.796, 6.022]
GP-MAE+COR	3.1712	[3.130, 3.212]	3.1032	[3.062, 3.145]	5.0804	[5.033, 5.127]	5.7368	[5.640, 5.834]
GP-MAE+COR+LR	3.1377	[3.094, 3.181]	3.0702	[3.027, 3.114]	5.0950	[5.054, 5.136]	5.7474	[5.679, 5.816]

remaining parameters are set to the defaults of ECJ package that our software implementation is built upon [9].

Our goal is to synthesize a model with possibly low Mean Absolute Error (MAE), i.e.,

$$f_{MAE} = \frac{1}{|I|} \sum_{i \in I} |\hat{y}_i - y_i|,$$

where y_i is the actual (observed) value of the heat flux, $\hat{y}_i$ is the output of a model, and I is the set of indices of samples (observations) of interest. f_{MAE} is thus our *objective performance measure*, and we use it for reporting throughout this paper. However, as we argued elsewhere [8], in general there is no rationale to claim that an objective performance is necessarily the best *search driver*, i.e., the best means to navigate the search process. The experimental evidence we gathered in the cited work actually suggests the opposite: in many cases, an alternative search driver may prove more effective at guiding an evolutionary search process (program/model synthesis in particular) towards the well-performing solutions.

There is a multitude of alternative search drivers that might be used for the heat flux problem. In our earlier studies, we compared a range of them when modeling the global temperature anomaly [14]. The overall conclusion was the superior efficiency of the Pearson correlation coefficient $\phi(\hat{y}, y)$. Following that past work, we define the (minimized) fitness as

$$f_{COR} = 1 - |\phi(\hat{y}, y)|.$$

f_{COR}, when applied as a fitness function in GP runs, was in most cases able to find models that were better (in terms of MAE) than the models evolved by GP driven by f_{MAE}. We attribute this interesting result to the fact that f_{COR} 'seeks' *any* linear dependency between the variables in question, and will for instance reward a model that captures the temporal dynamics of the dependent variable but diverges in amplitude (or in offset) from the desired values. A linear transformation of the output of such a model may form a very good predictor. $\phi(\hat{y}, y)$ implicitly performs such a transformation by standardizing the variables in question, i.e., subtracting the average and dividing by standard deviation. By this token, for any $\hat{y}$ and y and any $\alpha, \beta \in \mathbb{R}$, $\alpha \neq 0$ it holds $\phi(\hat{y}, y) = \phi(\alpha\hat{y} + \beta, y)$. f_{MAE}, to the contrary, puts an emphasis on the absolute differences between the actual and predicted values, and may overlook such linearly-related models.

Given a symbolic regression problem, GP can usually quickly find a model that coarsely captures the dependencies between the variables. However, further progress may be difficult, because significant improvement may require fundamental overhaul of a model, which may be hard to achieve, especially with a single application of a search operator. This in turn may lead to premature convergence, symptomized by a population becoming dominated by very similar or even identical models. To mitigate this problem, several population diversification methods have been proposed in the past. Here, we turn the original model synthesis process from a single-objective into a multi-objective one. Technically, we evaluate each model with respect to f_{MAE} *and* f_{COR}, and treat these two measures as search objectives to be used in parallel. To exploit the multi-aspect characteristic captured in these objectives without naively aggregating them into one scalar value, we employ the Non-domination Sorting Genetic Algorithm (NSGA-II [4]), a state-of-the art technique of multiobjective selection. To select the most promising candidate solutions (here: models), NSGA-II builds a Pareto-ranking of the combined current population and an *archive*, and uses tournament selection on Pareto ranks to select the parents solutions that give rise to the next generation of candidate solutions. Ties on Pareto ranks are resolved using a *sparsity* measure, which promotes the solutions that feature unique characteristics in terms of criteria. See [4] for the detailed coverage of NSGA-II.

6. THE EVOLVED MODELS

In this section, we report the aggregated statistics on particular approaches. Following the arguments made in the previous section, we designed four GP configurations:

1. GP-MAE: single-objective GP driven by f_{MAE},
2. GP-COR: single-objective GP driven by f_{COR},
3. GP-MAE+COR: two-objective GP driven by f_{MAE} and f_{COR},
4. GP-MAE+COR+LR: two-objective GP driven by f_{MAE} and f_{COR}, with modified selection of the best-of-run model.

Each GP variant works with a population of 5000 models (individuals, candidate solutions) and lasts for 50 generations. When a GP run terminates, we select from its final population the *best-of-run* model. For GP-MAE and GP-MAE+COR it is the model with the smallest f_{MAE}. For GP-COR, it is the model with the smallest f_{COR}. GP-MAE+COR+LR runs proceed exactly as GP-MAE+COR and diverge from it only in the way the best-of-run model is appointed. For each model in the Pareto front of the final population, we perform linear regression of the output of that model $\hat{y}$ onto the output variable y *for the training data*. Then we calculate f_{MAE} of the resulting compound

model, i.e., of $\alpha\hat{y}+\beta$, with α and β, also on the training set. The model with the smallest f_{MAE} calculated in this way is the best-of-run individual.

For each configuration, 30 independent evolutionary runs were performed. Table 2 presents the MAE of the best-of-run models averaged over the runs, accompanied with 0.95-confidence intervals. In the same table, we report also two baselines: the performance of the conventional model (2) and of multiple linear regression (LR).

The outcomes for particular datasets are substantially different, so in the following we discuss them separately. The interesting feature of all SHEBA-trained models is that they do not seem to overfit: the performance on the training and test part is very similar. In terms of averages, the test error is actually smaller than the training one for all methods. As this relationship holds also for the conventional model, which does not involve any training, we hypothesize that this is more due to the particular random division of the SHEBA dataset into the training and test part, rather than due to the merits of particular methods (note that overfitting does take place for the ISW dataset).

In terms of average error, LR fares the worst, indicating that the underlying phenomenon cannot be captured using linear dependencies. The nonlinear conventional model achieves a notably better performance. Nevertheless, the models synthesized by all variants of GP improve upon this result even further. As their upper confidence intervals are far from the performance of the conventional model, we may deem this difference statistically significant. The GP-COR method seems to perform the best on this dataset, reducing the prediction error by almost 20 percent on average, when compared to the conventional model. Nevertheless, the proximity of its confidence intervals to those of the other methods does not allow us find the differences between the GP variants statistically significant.

All GP-based methods offer a remarkably low variance of MAE. The leader in this respect is again GP-COR: its 0.083-wide confidence interval is less than 3 percent of the average. This suggest that GP can, despite inherent stochasticity, serve as a robust technique for modeling heat flux.

The results for the Antarctic ISW dataset present a less coherent picture. All models perform here much worse than on SHEBA, which is in part due to much larger variance of the observed heat flux for the ISW dataset ($\sigma_H = 11.83$, compared to 8.68 for SHEBA). Another factor is the smaller size of the ISW dataset (515 training observations vs. 3,004 for SHEBA). The conventional model commits the greatest error overall (we discuss this surprising underperformance in Section 7). LR fares better, but still worse than on SHEBA, which suggests that the ISW dataset is inherently more difficult to model. Despite this, GP is still capable to make better predictions.

For both datasets, using f_{COR} as a search driver is beneficial when compared to GP-MAE. This is confirmed also in terms of the overall-best models: the SHEBA model with the lowest training MAE (of all 30 evolutionary runs) of 2.843 has been found by GP-COR; for ISW, the overall-best model with f_{MAE}= 4.914 evolved in GP-MAE+COR. The use of linear regression in GP-MAE+COR+LR also reduces the error in comparison to GP-MAE+COR, albeit only slightly.

The disparity of MAE between the SHEBA and ISW datasets is also reflected in the sizes of the evolved models, which we report in Table 3. The models evolved for ISW are

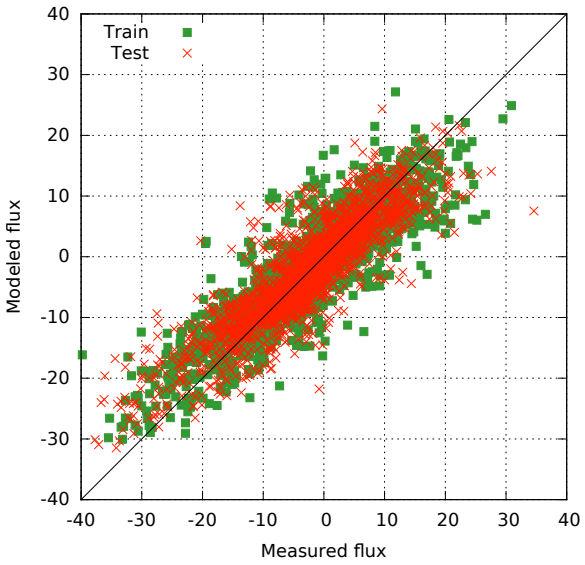

Figure 1: Best SHEBA model vs. measurements.

systematically more complex than those evolved for SHEBA. Apparently, in order to attain evolutionary advantage, the consecutive generations of good models in ISW runs needed to get more and more complex, and do so at a greater rate than for SHEBA (the GP runs for both datasets start from identical initial populations). This might have to some extent contributed to the overfitting on MAE (cf. Table 2). More importantly however, the models evolved using the bi-objective approach (GP-MAE+COR, GP-MAE+COR+LR) tend to produce, for both datasets, smaller models than those evolved by the single-objective GP-COR. The models in question are not smaller than those produced by GP-MAE, but GP-MAE is the least attractive configuration in terms of error. The GP-MAE+COR configuration offers a compromise between accuracy and complexity.

The models may seem large in absolute terms, but can be usually simplified. In Fig. 4, we present the simplified version of the overall smallest model, which happened to be evolved by GP-COR for the SHEBA dataset. The MAE of this model on the training and test set amounts to, respectively, 3.168 and 3.104.

7. COMPARISON WITH OBSERVATIONS AND CONVENTIONAL MODEL

In this section, we take a closer look at the characteristics of the evolved models and discuss them also from the meteorological perspective.

We first consider the model with the lowest MAE on the training set for the SHEBA dataset. This model comprises 188 nodes and commits MAE of 2.844 and 2.813 on the training and test set, respectively. In Fig. 1, we confront the observed heat flux (horizontal axis) with the predicted one (vertical axis), for both the training and the test part of this dataset. The plot reveals strong correlation of model's predictions with the observed data. The outliers are few and far between, so the model may be considered robust.

In Fig. 2, we present an analogous graph for the smallest SHEBA model, comprising 47 nodes and committing MAE of 3.168 and 3.104 on the training and test set, respectively.

Table 3: Complexity of the evolved best-of-run models, expressed as the number of nodes in expression trees, averaged over 30 evolutionary runs (with 0.95 confidence intervals).

	SHEBA dataset (Arctic)		*ISW dataset (Antarctic)*	
GP-MAE	97.27	[81.52, 113.02]	103.03	[89.35, 116.72]
GP-COR	137.77	[121.8, 153.74]	183.47	[160.59, 206.34]
GP-MAE+COR	101.37	[86.75, 115.98]	140.27	[122.55, 157.99]
GP-MAE+COR+LR	107.07	[94.56, 119.57]	144.83	[129.56, 160.11]

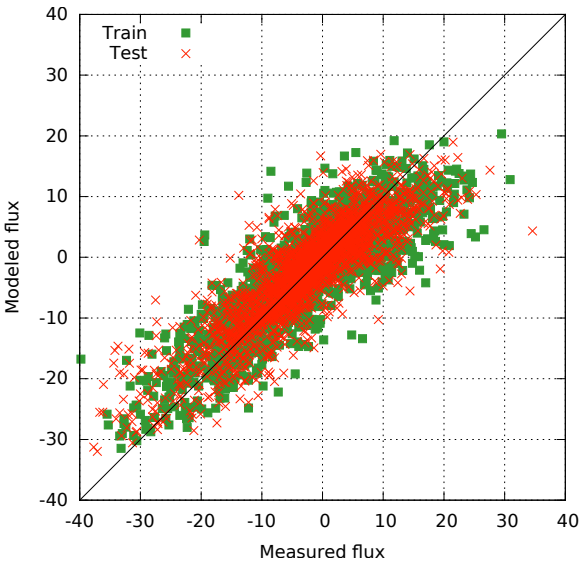

Figure 2: Smallest SHEBA model vs. measurements.

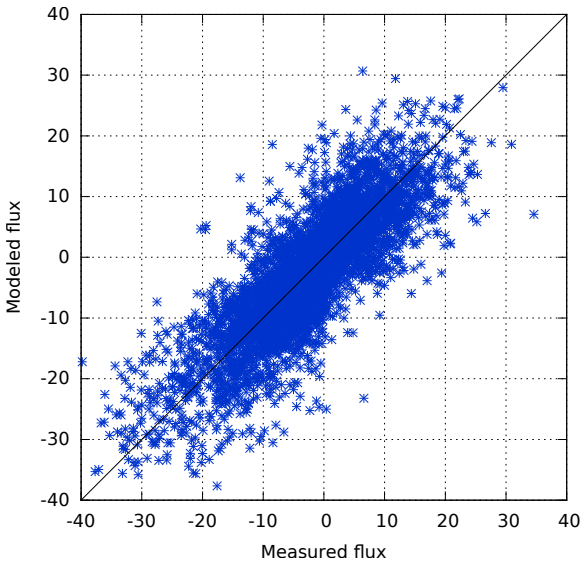

Figure 3: Conventional model vs. measurements (SHEBA).

The datapoints are more dispersed than in Fig. 1, but overall this model's behavior is also predictable.

For reference, Fig. 3 shows the same type of graph of the conventional model. Expectedly, the datapoints are even more dispersed than in Fig. 2, which reflects the overall worse MAE of this model.

A closer inspection of these graphs reveals an interesting difference between the GP models and the conventional model: the former tend to better at predicting negative flux, while the latter does not seem to show a significant preference in this respect.

Figures 6-8 present the analogous graphs for the ISW dataset: the lowest-MAE model (Fig. 6, MAE of 4.914 on training and of 5.840 on testing set), the smallest model (Fig. 7, MAE of 5.477 on training and of 6.016 on testing set), and the conventional model (Fig. 8). As expected, the points are more dispersed due to larger overall MAE. Interestingly, the tendency of producing better predictions for negative fluxes observed for SHEBA seems to occur here as well for the evolved models. The conventional model in Fig. 8 strongly underestimates the flux and occasionally exhibits tendency for predicting zero flux (notice the prominent horizontal agglomeration of observations). Both these phenomena can be explained by the context in which the ISW dataset was collected, which is however beyond the scope of this paper.

We attempt now to interpret the causality revealed by the smallest models evolved for both datasets. The smallest model for the SHEBA dataset, shown in Fig. 4, comprises 36 nodes after simplification. The dominant effect of the air-surface temperature difference (*TempDiff*) seems to be that surface warmer than the air favors an upward heat flux, in accordance with the conventional model in (2). However, the temperature difference appears in several places of the model, suggesting that the relationship between the heat flux and temperature difference is fairly complex. Both in (2) and this model, a strong wind (*WindSpeed*) favors a large heat flux, which is relevant as the wind shear is the primary source of turbulence in conditions when the surface is not strongly heated (which is always the case over sea ice). Net shortwave radiation (*NetShort*) tends to reduce the heat flux. The direct causal effect should be opposite, as net shortwave radiation heats the snow/ice surface favoring a heat flux from snow/ice to air. However, in SHEBA clouds warmed the snow/ice surface all the year except July and August, and their effect in increasing downward longwave radiation dominated over their effect in decreasing downward shortwave radiation [13]. Hence, under clear skies, with a lot of net shortwave radiation, the snow/ice surface was typically colder than under cloudy skies with less net shortwave radiation. Finally, a large specific humidity (*SpecHum*) is associated with a large heat flux. This is related to the above-mentioned cloud effects, as cloudy skies are associated with a large specific humidity and a warm snow/ice surface, favoring an upward heat flux.

The smallest model evolved for the ISW dataset is shown in Fig. 5 and comprises 44 nodes (60 nodes before simplification). In this model, the air-surface temperature difference and wind speed are both major factors, as in the conventional model in (2), but the complexity of their combined effect may suggest that (2) is too simplistic (we note that the temperature difference and wind speed also affect the static

stability that affects C_H in (2)). The role of net shortwave radiation is opposite to that in SHEBA. This is probably related to the fact that ISW drifted at lower latitudes, and therefore the surface temperature under clear skies, with a lot of solar radiation, is not necessarily lower than under cloudy skies.

It is noteworthy that, contrary to the conventional model, the discussed SHEBA and ISW models were not solely based on the air-surface temperature difference and wind speed, but also on the radiative fluxes, which might have helped reducing their errors.

8. DISCUSSION AND CONCLUSIONS

In this study, we demonstrated the feasibility of using GP to synthesize viable models of turbulent heat flux of sensible heat from experimental data. Many evolved models we produced here outperformed the method applied in the ECMWF operational weather forecasting model, which is considered the world's best numerical weather prediction model (see e.g. the forecast score statistics calculated by the World Meteorological Organization[2]). The best model commits MAE of 2.813 on the test part of the SHEBA dataset, i.e., 24 percent smaller than the error of the conventional model. The best model evolved for the ISW dataset has test-set MAE of 5.840, 51 percent smaller than the conventional model.

When it comes to GP-related aspects of this study, we brought substantial evidence that using an alternative 'search driver' (here: f_{COR}) as a fitness function in place of the objective that is normally employed to assess the models in a domain (f_{MAE}) can have essential positive impact on the outcome. This observation holds also for the multiobjective evolutionary approaches, which, though not necessarily outperform the single-objective methods on error, may improve the models with respect to other, 'non-functional' properties like complexity.

There are several directions in which this research can be taken further. This study abstracts from the temporal nature of the heat flux phenomenon; as a consequence, the models we evolved here are inherently incapable of capturing the cumulative characteristics of the underlying physics. A natural next step could be thus to extend the repertoire of input variables (which now reflect the state of the system at a given time instant) with extra variables that capture the history of the process.

Employing a model evolved by GP in a numerical weather prediction or climate model would open the door to many interesting possibilities. A more accurate parameterization for the surface heat flux should directly improve the accuracy of the simulated surface heat budget (1), surface temperature, and near-surface air temperature. Further, it should also indirectly improve the simulations of near-surface wind speed, air humidity, occurrence of fog, melt of snow and ice, and growth of sea and lake ice. These are essential issues in both numerical weather prediction and climate research, and particularly actual now when a dramatic loss of sea ice and terrestrial snow is going on in the Arctic.

Acknowledgments. The work of KS and TV was supported by the Academy of Finland through the CACSI project (contract 259537). KK acknowledges support from project 09/91/DSPB/0572.

[2]http://apps.ecmwf.int/wmolcdnv/scores/mean/850_t

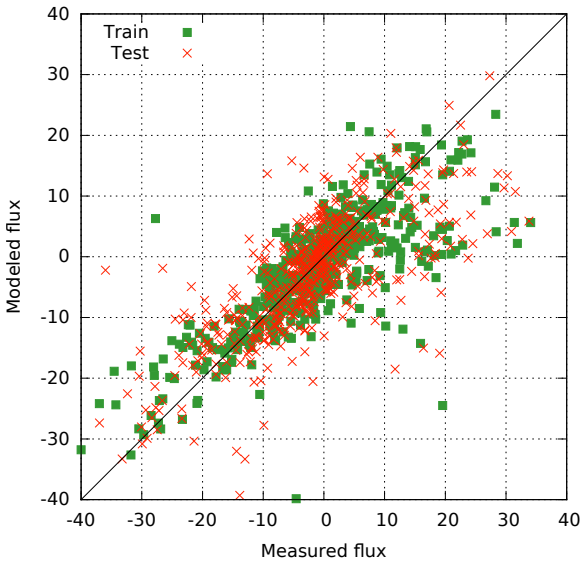

Figure 6: Best ISW model vs. measurements.

9. REFERENCES

[1] E. L. Andreas, R. E. Jordan, and A. P. Makshtas. Parameterizing turbulent exchange over sea ice: The Ice Station Weddell results. *Boundary-Layer Meteorology*, 114:439–460, 2005.

[2] Evgeny Atlaskin and Timo Vihma. Evaluation of NWP results for wintertime nocturnal boundary-layer temperatures over Europe and Finland. *Quarterly Journal of the Royal Meteorological Society*, 138(667):1440–1451, 2012.

[3] D. J. Cavalieri and C. L. Parkinson. Arctic sea ice variability and trends, 1979–2010. *The Cryosphere*, 6(4):881–889, 2012.

[4] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *Evolutionary Computation, IEEE Transactions on*, 6(2):182 –197, apr 2002.

[5] C. Derksen and R. Brown. Spring snow cover extent reductions in the 2008–2012 period exceeding climate model projections. *Geophysical Research Letters*, 39(19), 2012.

[6] R. Döscher, T. Vihma, and E. Maksimovich. Recent advances in understanding the Arctic climate system state and change from a sea ice perspective: a review. *Atmospheric Chemistry and Physics*, 14(24):13571–13600, 2014.

[7] John R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA, USA, 1992.

[8] Krzysztof Krawiec and Una-May O'Reilly. Behavioral programming: A broader and more detailed take on semantic GP. In *Proceeding of the sixteenth annual conference on Genetic and evolutionary computation conference*, GECCO '14, New York, NY, USA, 2014. ACM.

[9] Sean Luke. ECJ evolutionary computation system, 2002. (http://cs.gmu.edu/ eclab/projects/ecj/).

[10] L. Mahrt. Stably stratified boundary layer. In J. P. Holton, J. A. Curry, and J. Pyle, editors, *Encyclopedia*

$$-TempDiff * 0.893^{TempDiff} * [\frac{WindSpeed}{0.798^{TempDiff}} - TempDiff + 0.0126 * (SpecHum - NetShort + 0.405) * (TempDiff - 0.0146 * NetLong * 0.636^{TempDiff})] * 1.953 + 0.39$$

Figure 4: The smallest best-of-run heat flux model evolved for the SHEBA dataset (originally composed of 47 nodes, manually simplified to 36 nodes). This model attains MAE of 3.168 on the training set and 3.104 on the test set.

$$[2 * (WindSpeed + TempDiff - WindSpeed * TempDiff) + 10^{-6} * LongOut^2 * WindSpeed * (NetShort + NetLong) + \sqrt{\frac{WindSpeed}{ShortOut+87.45}} * (NetLong + 0.001 * LongOut * WindSpeed * (NetShort + NetLong))] * 0.405 + WindSpeed$$

Figure 5: The smallest best-of-run heat flux model evolved for the ISW dataset (originally composed of 60 nodes, manually simplified to 44 nodes). This model attains MAE of 5.476 on the training set and 6.016 on the test set.

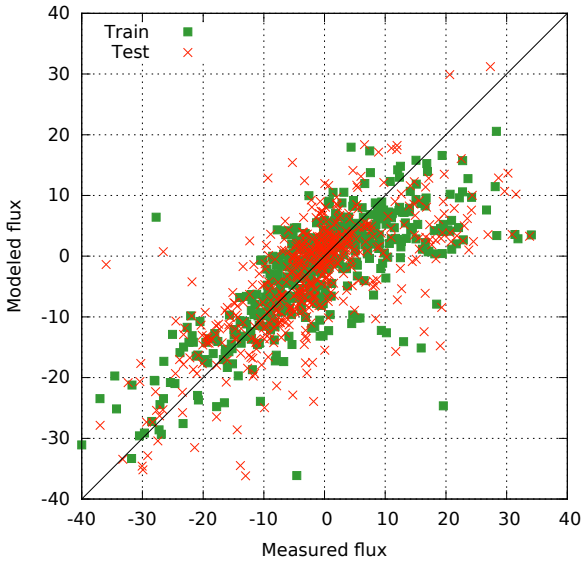

Figure 7: Smallest ISW model vs. measurements.

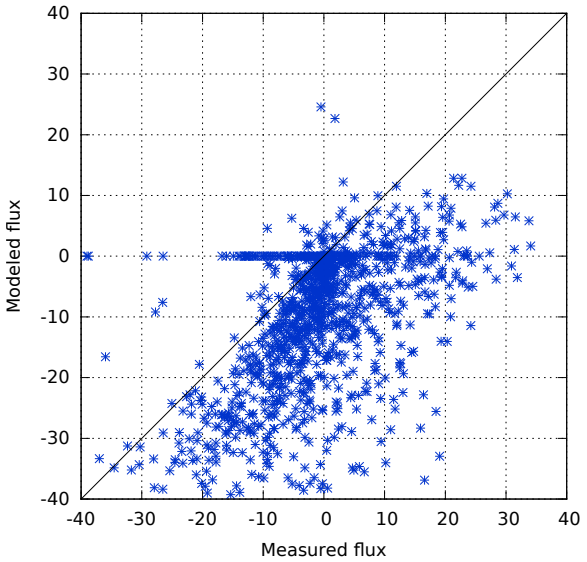

Figure 8: Conventional model vs. measurements (ISW).

of Atmospheric Sciences, pages 298–305. Academic Press, London, 2002.

[11] A. S. Monin and A. M. Obukhov. Basic laws of turbulent mixing in the surface layer of the atmosphere (in Russian). *Tr. Akad. Nauk SSSR Geofiz. Inst.*, 24:163–187, 1954.

[12] P. Ola G. Persson, Christopher W. Fairall, Edgar L. Andreas, Peter S. Guest, and Donald K. Perovich. Measurements near the Atmospheric Surface Flux Group tower at SHEBA: Near-surface conditions and surface energy budget. *Journal of Geophysical Research: Oceans*, 107(C10):SHE 21–1–SHE 21–35, 2002.

[13] M. D. Shupe and J. M. Intrieri. Cloud radiative forcing of the Arctic surface: the influence of cloud properties, surface albedo, and solar zenith angle. *Journal of Climate*, 17:616–628, 2004.

[14] Karolina Stanislawska, Krzysztof Krawiec, and Zbigniew W. Kundzewicz. Modelling global temperature changes with genetic programming. *Computer & Mathematics with Applications*, 64(12):3717–3728, 2012.

[15] G. Svensson, A.A.M. Holtslag, V. Kumar, T. Mauritsen, G.J. Steeneveld, W.M. Angevine, E. Bazile, A. Beljaars, E.I.F. de Bruijn, A. Cheng, L. Conangla, J. Cuxart, M. Ek, M.J. Falk, F. Freedman, H. Kitagawa, V.E. Larson, A. Lock, J. Mailhot, V. Masson, S. Park, J. Pleim, S. Söderberg, W. Weng, and M. Zampieri. Evaluation of the diurnal cycle in the atmospheric boundary layer over land as represented by a variety of single-column models: the second GABLS experiment. *Boundary-Layer Meteorology*, 140(2):177–206, 2011.

[16] M. Tjernström, M. Zagar, G. Svensson, J.J. Cassano, S. Pfeifer, A. Rinke, K. Wyser, K. Dethloff, C. Jones, T. Semmler, and M. Shaw. Modelling the Arctic boundary layer: an evaluation of six ARCMIP regional-scale models using data from the SHEBA project. *Boundary-Layer Meteorology*, 117:337–381, 2005.

[17] T. Vihma, R. Pirazzini, I. Fer, I. A. Renfrew, J. Sedlar, M. Tjernström, C. Lüpkes, T. Nygård, D. Notz, J. Weiss, D. Marsan, B. Cheng, G. Birnbaum, S. Gerland, D. Chechin, and J. C. Gascard. Advances in understanding and parameterization of small-scale physical processes in the marine Arctic climate system: a review. *Atmospheric Chemistry and Physics*, 14(17):9403–9450, 2014.