Computers & Geosciences 106 (2017) 139–149

Contents lists available at ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Research paper

Application of evolutionary computation on ensemble forecast of quantitative precipitation

Amanda S. Dufek[a,*], Douglas A. Augusto[b], Pedro L.S. Dias[c], Helio J.C. Barbosa[a,d]

[a] National Laboratory for Scientific Computing, 333 Getúlio Vargas Avenue, Quitandinha, Petrópolis, RJ 25651-075, Brazil
[b] Oswaldo Cruz Foundation, Rio de Janeiro, RJ, Brazil
[c] Institute of Astronomy, Geophysics and Atmospheric Sciences, University of São Paulo, SP, Brazil
[d] Federal University of Juiz de Fora, Juiz de Fora, MG, Brazil

ARTICLE INFO

Keywords:
Ensemble weather forecast
Quantitative precipitation
Evolutionary computation
Genetic programming

ABSTRACT

An evolutionary computation algorithm known as genetic programming (GP) has been explored as an alternative tool for improving the ensemble forecast of 24-h accumulated precipitation. Three GP versions and six ensembles' languages were applied to several real-world datasets over southern, southeastern and central Brazil during the rainy period from October to February of 2008–2013. According to the results, the GP algorithms performed better than two traditional statistical techniques, with errors 27–57% lower than simple ensemble mean and the MASTER super model ensemble system. In addition, the results revealed that GP algorithms outperformed the best individual forecasts, reaching an improvement of 34–42%. On the other hand, the GP algorithms had a similar performance with respect to each other and to the Bayesian model averaging, but the former are far more versatile techniques. Although the results for the six ensembles' languages are almost indistinguishable, our most complex linear language turned out to be the best overall proposal. Moreover, some meteorological attributes, including the weather patterns over Brazil, seem to play an important role in the prediction of daily rainfall amount.

1. Introduction

The main goal of this paper is to propose a new approach based on genetic programming algorithms to create more accurate deterministic ensemble forecasts (DEF) of 24-h accumulated precipitation. This goal is motivated by the importance of an accurate and reliable quantitative precipitation forecast (QPF) for the strategic planning of several socio-economic sectors (such as agricultural production, hydropower generation, water availability for public consumption, and flood and landslide control), as well as by the difficulty in forecasting quantitative precipitation and by the limitations of the current methods for postprocessing ensembles. The traditional statistical techniques (such as model output statistics (MOS; Glahn and Lowry, 1972), MASTER super model ensemble system (MSMES; Silva Dias et al., 2006), and Bayesian model averaging (BMA; Raftery et al., 2005)) have worked well for variables such as temperature and geopotential height. However, these approaches lead to unsatisfactory results for QPF, perhaps because the distribution of precipitation is far from normal (usually gamma distribution), or due to the complexity of the processes involved, or because of its high spatial, temporal and frequency variability.

Genetic programming (GP) is an evolutionary algorithm, which is inspired by genetics and Darwinian evolution. GP was introduced by Koza (1992) in the early 1990s, due to its ability to learn implicit relationships in observed data and to express them automatically in a symbolic mathematical manner. Furthermore, GP is a supervised machine learning technique that has been able to solve complex optimization problems which cannot feasibly be solved directly or rigorously in real-world applications. Gene-expression programming (GEP) (Ferreira, 2001), grammar-based GP (GGP) (Whigham, 1995) and grammatical evolution (GE) (Ryan et al., 1998) are specializations of the canonical GP, with the last two having the advantage of evolving syntactically correct solutions in an arbitrary language described by a grammar.

In contrast to traditional statistical approaches, evolutionary algorithms do not require prior knowledge about the statistical distribution of the data, nor do they need to explicitly assume a model form. Moreover, evolutionary algorithms usually test many solutions instead of continually trying to improve a single one, and can also automatically capture complex interactions among input and output variables in a system. Additionally, the ability of traditional statistical techniques to deal with non-linear problems is limited, whereas for the evolutionary

* Corresponding author.
E-mail address: amandasd@lncc.br (A.S. Dufek).

http://dx.doi.org/10.1016/j.cageo.2017.06.011
Received 19 July 2016; Received in revised form 31 January 2017; Accepted 7 June 2017
Available online 10 June 2017
0098-3004/

algorithms it is very satisfactory.

Until recently, only a few papers focused on applying GP algorithms in Hydrology, Meteorology and Water Resources (Omolbani et al., 2010). For the ensemble forecast problem, Bakhshaii and Stull (2009) proposed the use of GEP to form linear or non-linear combinations of numerical weather predictions (NWP). The authors applied GEP to produce short-range DEFs of 24-h accumulated precipitation at 24 stations in mountainous southwestern Canada during the two fall–spring rainy seasons of October 2003–March 2005, using an eleven-member multimodel multigrid-size ensemble. The GEP DEFs obtained superior performance relative to simple ensemble means for about half of the mountain weather stations tested. Roebber (2010) focused on the production of consensus 24-h forecasts for minimum temperature at a site in Ohio derived from evolutionary programming (EP). The resulting deterministic forecasts' improvement relative to MOS was nearly 27%. Roebber (2015a) (2015b) extended this work to investigate probabilistic as well as deterministic forecasts of minimum temperature, which were superior to those obtained from operational ensembles and MOS.

Roebber's papers are concerned with generating ensemble of EP solutions, whereas here we are interested in optimizing a combination of NWP ensemble members as in Bakhshaii and Stull (2009). Two important differences between the purpose of this paper and that of Bakhshaii and Stull (2009) are: (i) the use of grammar-based GP instead of GEP, a non-grammatical approach, and (ii) the inclusion of other potential predictors, such as the major weather patterns over Brazil, in addition to NWP models. Although in Roebber, (2010, 2015a, 2015b) the author introduces specialist's domain knowledge into the programs' language, this is not achieved through a formal grammar as in our work. Furthermore, QPF for regions of Brazil is considered a harder problem than minimum temperature forecasting, as addressed by Roebber, (2010, 2015a, 2015b), due to the more complex processes associated with tropical and subtropical convection.

The current paper is an extension to the previous work (Dufek et al., 2013) in which the feasibility of the GE algorithm to deal with the problem of ensemble forecast of rainfall amount was evaluated on three artificial datasets comprising known relationships between three hypothetical meteorological models and two weather patterns. Now, three GP versions are applied to postprocessing short-range ensemble forecast of daily rainfall amount for several real-world datasets. Furthermore, other meteorological information are incorporated into the grammars in addition to weather patterns.

The main contributions of this paper consist of (i) creating deterministic ensemble 24- and 72-h forecasts of 24-h accumulated precipitation based on GGP and GE algorithms for 317 locations in southern, southeastern and central Brazil during the rainy period from October to February of 2008–2013; (ii) comparing in terms of accuracy the DEFs of quantitative precipitation via GP algorithms with those obtained from three traditional statistical techniques: simple ensemble mean, MSMES, and BMA, and also with the best forecast in the ensemble; (iii) the development and study of six different ensemble forecast grammars to represent the possible solutions to the ensemble forecast problem; (iv) an investigation into the non-linearity of the phenomenon; (v) providing some meteorological information as input attributes in order to enrich the GP forecasting model; (vi) an investigation into the influence of the four major weather patterns in Brazil on the precipitation skill of NWP models; (vii) extracting knowledge from the resulting best solutions, such as the relationships between the input attributes and the occurrence of rainfall, and the classification of the meteorological attributes in order of importance in the ensemble postprocessing.

The frequently used abbreviations are listed in Table 1 in order to facilitate the reading of the paper.

Table 1
List of frequently used abbreviations in this paper.

Abbreviation	Description
BESTFCST	best ensemble member
BMA	Bayesian model averaging—a performance-based weighted ensemble mean (see Section 3.2.1)
BMA-P	pattern-based BMA (see Section 3.2.1)
DEF	deterministic ensemble forecast
GE	grammatical evolution
GE_1	grammatical evolution with simultaneous approach (see Section 3.2.2 for more details)
GE_2	grammatical evolution with decoupled approach (see Section 3.2.2)
GGP	grammar-based genetic programming
GGP_2	grammar-based genetic programming with decoupled approach (see Section 3.2.2)
GP	genetic programming
MAE	mean absolute error
MSMES	MASTER super model ensemble system—a performance-based weighted ensemble mean (see Section 3.2.1)
MSMES-P	pattern-based MSMES (see Section 3.2.1)
NWP	numerical weather predictions
QPF	quantitative precipitation forecast
SM	simple mean—an equally-weighted ensemble mean (see Section 3.2.1)

2. Genetic programming

GP is one of the main areas of evolutionary computation, first devised by Cramer (1985) and greatly developed by Koza (1992). GP is a stochastic optimization technique based on Darwin's theory of evolution by natural selection that evolves a population of computer programs, usually expressed as syntax trees. Whigham (1995) introduced the grammar-based GP (GGP) in order to evolve syntactically correct computer programs in an arbitrary language described by a grammar. Grammatical evolution (GE) (Ryan et al., 1998) is a variation of GGP in which the computer programs are encoded in linear structures instead of tree-based data structures typical of GP and GGP.

Next, we give a brief overview of the concept of GP (Eiben and Smith, 2003), whose algorithm is outlined in Algorithm 1.

Algorithm 1: General scheme of genetic programming in pseudo-code.

```
Generate a random initial population
Evaluate all the individuals
while Stopping criterion is not satisfied do
    Select parents from the current population
    Apply genetic operators: crossover and mutation to the previously selected parents
    Evaluate the resulting offspring
    Select individuals for the next generation
```

GP algorithm is population based, i.e. it processes a whole collection of candidate solutions simultaneously. Each candidate solution—also called individual or computer program—is evaluated according to some fitness function which assigns a quality measure to the individuals. Based on this fitness, some of the candidates are stochastically selected from the current population to seed the next generation by applying genetic operators to them. The selection operator ensures a bias towards fitter individuals. Nevertheless, it also allows for the occasional selection of less-fit individuals, since otherwise the whole search could become too "greedy" and get stuck in a local optimum. Two of the most important genetic operators are crossover and mutation. Similarly to selection operators, crossover and mutation are stochastic operators. Crossover merges information from two or more selected candidates—the so-called parents—to generate one or more new candidates—the offspring. Mutation causes a small undir-

ected change in one candidate, resulting in a new one. Crossover essentially exploits inherent possibilities in the population, while mutation creates random diversity in the population by exploiting newly created genetic material. The evolutionary process follows with the replacement strategy, in which the old and new individuals compete for a place in the next generation. The evolutionary process of evaluation, selection, genetic operators, and replacement is iterated until a stopping criterion is satisfied, pressing forward the improvement of the population's fitness at each iteration.

Population-based model and stochastic operators are the two main characteristics of GP that make it a very robust technique. In addition, it is a relatively straightforward algorithm in terms of concept and implementation, and also exhibits inherent parallelism. Another advantage of GP is the minimal requirement of domain knowledge, besides operating well on discontinuous search spaces. One of its major advantages over other techniques is its capacity of evolving human-interpretable solutions of potentially unbounded complexity. An additional advantage of grammar-based GP is the possibility of introducing specialist knowledge into the grammar in the hope of finding a better-quality, syntactically correct solution in a shorter period of time.

2.1. Input attributes

Next, we describe several input attributes used by the grammars as operands of the language. Below is the (i) to (xv) list of input attributes. Table 2 provides a short description of them for a quick consultation. Input attributes are highlighted in bold throughout the text.

(i) **O(1day):** Rainfall amount observed on the day before the forecast date.

(ii) **O(2day):** Rainfall amount observed on the day preceding the day before the forecast date.

(iii) **O(mean):** Mean of the observed daily rainfall amount calculated from an 11-day window centered on each calendar day in the base period 1998–2013. An 11-day window is chosen to yield a total sample size of 16 years × 11 days = 176 days for each calendar day.

(iv) **O(lag1+)**, **O(lag1-)**, **O(lag2+)**, **O(lag2-)**, **O(lag3+)** and **O(lag3-):** Let an observational gridded dataset be given by the following four daily mean variables: temperature, zonal and meridional wind speed at 850hPa, and specific humidity at 750hPa drawn from the NCEP/CFSR dataset (Saha et al., 2010), in the domain between 50°S–10°N and 82–34.5°W, during the period from October to February of 2008–2013. For each location, its one-point correlation map with lag-L, $L = 1, 2, 3$, was constructed at 2.5° spatial resolution. Displayed on the one-point correlation map with lag-L are the contours of Pearson correlations between the data at 25 latitude × 20 longitude = 500 grid points shifted by L days and time-unlagged data at the location of interest (Wilks, 2006, chap. 3). There are as many maps as locations. For each one-point correlation map, the observed 24-h accumulated precipitation from the points with the largest negative and positive correlations were extracted and called, respectively, **O(lag1+)** and **O(lag1-)** for $L = 1$. Since we are interested in teleconnection patterns between the target location and remote regions, the correlation values from a 5 × 5 square grid centered at the target location were excluded from the analysis.

(v) $\mathbf{M}_i$: QPF provided by the member i.

(vi) **M(mean):** Average of QPF ensemble members, no bias correction.

(vii) **M(min):** Minimum value among the ensemble member forecasts.

(viii) **M(max):** Maximum value among the ensemble member forecasts.

(ix) **M(std):** Standard deviation of ensemble member forecasts.

(x) **rain:** A binary attribute defined based on the following criterion: **M(mean)** above 1 mm assumes 1; otherwise 0.

(xi) **BMA:** BMA DEF of 24-h accumulated precipitation. **BMA** (in bold typeface) is used to indicate the attribute, while BMA (in regular typeface) refers to the technique.

(xii) **P:** The traditional k-means clustering algorithm was applied to identify the four major weather patterns over Brazil. The input data to the k-means algorithm are seven daily fields of the NCEP/CFSR dataset. For a domain between 40°S–0° and 67–19.5°W, the fields include zonal and meridional gradients of specific humidity at 750hPa and temperature at 850hPa, and vorticity at 850, 500 and 200hPa during the period from October to February of 1979–2013. The k-means algorithm clusters the dataset, with 5 193 registers available, into four groups based upon the N-dimensional Euclidean distance, with $N = 7$ fields × 17 latitude × 20 longitude = 2 380. A typical weather pattern of Brazil was defined as composites of atmospheric fields belonging to a given group. At the end, each register is classified under one of four predefined weather patterns. Three of the four patterns are associated with the propagation of frontal systems from southwest to northeast, two of which indicate the configuration of a typical South Atlantic Continental Convergence Zone episode (Carvalho et al., 2004). In the last pattern, the presence of a Mesoscale Convective System over Rio Grande do Sul, Uruguay and parts of Argentina is highlighted (Velasco and Fritsch, 1987).

(xiii) **O(P):** Rainfall from the composites of the observed rainfall amount fields belonging to each group, i.e. rainfall from cluster centroids.

(xiv) **pattern_change:** A binary attribute based on the following criterion: if the day before the forecast date and the target prediction day share the same weather pattern, then assumes 0; otherwise 1.

(xv) **K**, **TT** and **SWEAT:** Atmospheric instability indices calculated from the NCEP/CFSR dataset. The indices consist of empirical measures derived from several kinematic and thermodynamic considerations and are used as indicators of summer-time convective rainfall. It is worth mentioning that we are not forecasting severe storm events, and the indices are used only to capture some characteristic of the atmosphere, such as the existence of a cold, dry air layer at medium-levels overlying or not a layer of warm and moist air at low-levels. Other indices,

Table 2
Short description of input attributes.

Abbreviation	Description
O(1day)	rainfall amount observed on the day before the forecast date
O(2day)	rainfall amount observed on the day preceding the day before the forecast date
O(mean)	mean of the observed daily rainfall amount
O(lag$L\pm$)	observed daily rainfall amount from the points with the largest negative and positive correlations with lag-L, $L = 1, 2, 3$
$\mathbf{M}_i$	daily rainfall amount provided by the member i
M(mean)	mean of ensemble member forecasts
M(min)	minimum value among the ensemble member forecasts
M(max)	maximum value among the ensemble member forecasts
M(std)	standard deviation of ensemble member forecasts
rain	if **M(mean)**> 1 mm, then assumes 1; otherwise 0
BMA	Bayesian model averaging deterministic ensemble forecast of daily rainfall amount
P	four major weather patterns over Brazil
O(P)	rainfall from the composites of the observed rainfall amount fields belonging to each weather pattern
pattern_change	if the weather pattern changes, then assumes 1; otherwise 0
K, **TT** and **SWEAT**	atmospheric instability indices

such as convective available potential energy, convective inhibition, and vertical wind shear, should be tested in the future.

2.2. Ensembles' languages

A grammar defines—depending on the desired level of complexity—what does and what does not characterize an ensemble. In other words, it basically specifies how the NWP models are combined (linearly or non-linearly), and which input attributes are allowed.

Six grammars were here designed to tackle the problem of DEF of 24-h accumulated precipitation via GP algorithms, they are: **L**, **LP**, **LA**, **NL**, **NLP** and **NLA**, where the first three grammars refer to linear grammars, and the last three refer to non-linear grammars. The linear grammars only allow linear combinations of the NWP models, while the non-linear grammars allow linear and non-linear combinations of the NWP models. The term "A" indicates that all the input attributes described in the previous section were incorporated into the grammar by adding new rules that specify the possible semantic relations among the new attributes, and the other components of the language; the term "P" indicates that only the effects introduced by the weather patterns **P** can be taken into account by the language; and the absence of both terms "A" and "P" indicates that the input attributes are only the ensemble members' QPFs.

Since Greybush et al. (2008) and Espinosa (2011) have emphasized the importance of including weather pattern information in the ensemble postprocessing, the **P** attribute received special attention through the grammars **LP** and **NLP**. The **NLA** grammar is our most powerful—and complex—grammar, since it is an extension to the **NLP** grammar that includes new input attributes. The four grammars: **L**, **LP**, **NL** and **NLP** are the same as employed in Dufek et al. (2013), except for some mathematical, logical and relational operators; a detailed description of them can be found therein.

Table 3 provides a short description of the six ensemble forecast grammars for a quick consultation. Grammars are highlighted in bold throughout the text.

3. Experiments

3.1. Data

Daily rainfall amount predicted by several NWP models for 317 locations in the domain between 32.8°S–14.8°S and 57.8°W–39.8°W, during the period from October to February of 2008–2013, came from the Center for Weather Forecast and Climate Studies (CPTEC) of the Brazilian National Institute for Space Research (INPE). The corresponding observed values of daily rainfall amount were derived from a higher quality gridded dataset at a spatial resolution of 0.25° (Rozante et al., 2010). The location of the 317 locations is shown in Fig. 1.

The selection of a subset of available NWP models was based on the lowest mean absolute error of the MSMES DEF achieved among all[1] the possible subsets with at least four members. This choice of minimum number of members was made somewhat arbitrarily. Although the maximum number of members had not been established in advance, it did not exceed eight members. The choice of MSMES is justified by its low computational cost and its wide use at meteorological centers of Brazil. The subset varied according to the location and the forecast range. The fourteen NWP models selected as ensemble members are listed in Table 4, which also includes their spatial resolution and relative frequency of selection as ensemble members for the 24- and 72-h forecasts in the 317 locations. CPTEC/INPE is responsible for running all the NWP models and making their results available, except the RAMSC model which is the responsibility of MASTER/USP.

[1] To give an idea, there are roughly $2^{|M|}$ subsets to be evaluated, where $|M|$ is the number of available NWP models.

Table 3
Short description of the six ensemble forecast grammars.

Abbreviation	Description
L	linear grammar
LP	linear grammar that includes the weather patterns
LA	linear grammar that includes all the input attributes described in Section 2.1
NL	non-linear grammar
NLP	non-linear grammar that includes the weather patterns
NLA	non-linear grammar that includes all the input attributes described in Section 2.1

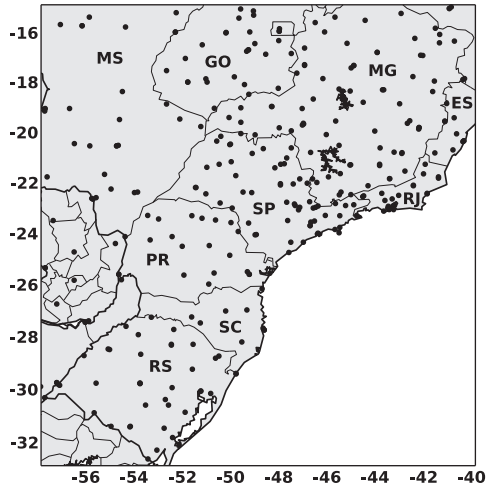

Fig. 1. Location of the 317 locations over southern, southeastern and central Brazil, along with parts of Uruguay, Argentina, Paraguay and Bolivia. The Brazilian states include: Mato Grosso do Sul (MS), Goiás (GO), Minas Gerais (MG), Espírito Santo (ES), São Paulo (SP), Rio de Janeiro (RJ), Paraná (PR), Santa Catarina (SC) and Rio Grande do Sul (RS).

3.2. Parameters and methodology

3.2.1. Traditional statistical techniques

Three traditional statistical techniques: simple ensemble mean, MSMES (Silva Dias et al., 2006) and BMA (Raftery et al., 2005; Sloughter et al., 2007) were used in order to postprocess short-range ensemble QPFs for the 317 locations. The simple ensemble mean is the simplest and most widely used technique for postprocessing ensemble forecasts. An improvement over the equally-weighted mean is the performance-based weighted mean, which takes into account the relative performance of ensemble members. The MSMES and BMA are examples of such a scheme. The three statistical techniques were used for comparison with the GP approaches.

The dependent variable is the observed 24-h accumulated precipitation at any one location. The predictors are the ensemble members' QPFs, referred to as $\mathbf{M}_i$, $i = 1,\ldots,m$, where m is the number of ensemble members (Section 2.1). The mean absolute error (MAE) was used as an accuracy measure for DEF of 24-h accumulated precipitation via the three statistical techniques. The DEFs were computed for each forecast range at each location.

For MSMES and BMA, the two different ways to define the training period are the same as applied in Dufek et al. (2013), and a description of them is derived from there as follows. The first one refers to the usual definition, whose training period consists of the immediately preceding t days. The second one is the pattern-based definition—hereafter referred to as MSMES-P and BMA-P—whose preceding training period is based on the similarity of weather patterns, and is consequently temporally noncontiguous: if P is the pattern of the target prediction day, then the training days will be the t previous ones that share the same P. From the viewpoint of practical applications, and

Table 4
The fourteen NWP models selected as ensemble members: ID, spatial resolution (km) and relative frequency (%) of selection as ensemble members for the 24- and 72-h forecasts in the 317 locations.

ID	Spatial Resolution	Relative Frequency (%)	
	(km)	24-h	72-h
GP213	50	68.1	36.9
CPTEC	100	19.9	19.9
SFENM	100	69.1	60.9
SFAVN	100	24.9	18.3
ACOPL	200	67.8	62.5
T299x	44	28.7	14.2
RPSAS	40	19.9	19.9
ETA20	20	19.9	19.9
ETAcr	40	33.1	37.9
ETAm1	40	20.2	29.7
ETAm2	40	22.7	23.7
ETAm3	40	13.2	22.1
ETAm4	40	19.2	22.4
RAMSC	25	7.6	69.7

Source: http://intercomparacaodemodelos.cptec.inpe.br/phps.

guided by previous works (Raftery et al., 2005; Eckel and Mass, 2005; Greybush et al., 2008), we have used a training period of t = 15 days for all the statistical techniques.

3.2.2. Genetic programming technique

Two GP approaches to address the problem of DEF of 24-h accumulated precipitation were proposed. The first approach—called a decoupled approach—decouples the DEF problem into smaller problems based on two main steps. In the first step, the GP algorithm is used for postprocessing each ensemble member separately. The next step consists of combining the resulting GP-corrected ensemble members into a more accurate single forecast via the GP algorithm in order to minimize the differences between ensemble forecast and observation. In the second approach—called a simultaneous approach—the GP algorithm combines the ensemble member forecasts while simultaneously correcting each of them.

A simple and straightforward version of the GE algorithm was implemented in the C programming language. It uses a standard binary code representation, one-point crossover and bit-flipping mutation operators, and adopts the policy of eliminating invalid non-decoding individuals from the evolutionary process. The generational replacement scheme with elitism and the tournament selection strategy were applied in the evolutionary search. In addition to GE, the GGP was implemented using the *EpochX* framework[2] (Otero et al., 2012). The implementation uses the standard subtree operators of mutation and crossover, ramped half-and-half tree initialization, generational replacement scheme with elitism, and tournament selection strategy.

The parameters were assigned empirically and remained constant throughout the experiments. GE and GGP were carried out with the following parameters: population of 2 000 programs, crossover rate of 90%, elitism of 1 program, and tournament size of 3 programs. The stopping criterion is given by a maximum number of generations set to 1 000 for the GE and GGP at the decoupled approach. For the simultaneous strategy, the maximum number of generations was set to 1 000+1 000×m, where m is the number of ensemble members, such that the computational effort is the same as that required for the decoupled strategy. The GE parameters were: per-bit mutation rate of 0.25%, genome size of 2 000 bits, gene size of 8 bits, ephemeral constants (Augusto et al., 2011) in the interval [−16,+16] (for **L**, **LP**, **NL** and **NLP** grammars) and [0,300] (for **LA** and **NLA** grammars), and numeric precision of 16 bits. Finally, the GGP parameters were: mutation rate of 10%, maximum initial tree depth of 8 nodes, maximum tree depth of 14 nodes, and numerical constants generated by the digit concatenation approach (O'Neill et al., 2003) in the interval [0,9.999].

Ten independent runs of each version (GE and GGP with simultaneous and decoupled approaches) and grammar (**L**, **LP**, **LA**, **NL**, **NLP** and **NLA**) of the evolutionary algorithm were performed for each forecast range (24- and 72-h) at each location (317 in total), yielding different solutions at each run. Independent runs means a different seed of the pseudo-random number generator, and random data distribution for the training and test sets for each one. The training set was composed by 90% of the instances (about 450 instances) while the test set consisted of the remaining 10% of the instances (about 50 instances). The fitness function was defined as the MAE in a given training set. The sample median from ten MAEs relative to training and test sets was calculated. Due to the high computational cost required by the GGP written in Java language, only the following experiments were performed: ensemble 72-h QPFs given by the GGP with decoupled approach for the **L**, **LP**, **NL** and **NLP** grammars.

The dependent variable is the daily rainfall amount observed at any one location. The predictors are all or some of the input attributes described in Section 2.1, depending on the grammar. It is worth noting that the **LA** and **NLA** grammars are only used by the GE with simultaneous approach.

3.3. Results

3.3.1. General analysis

Fig. 2 shows the box plots of the MAE of the deterministic ensemble 24-h QPF achieved through GE with simultaneous approach (GE_1) for the six grammars (**L**, **LP**, **LA**, **NL**, **NLP** and **NLA**) and GE with decoupled approach (GE_2) for the four grammars (**L**, **LP**, **NL** and **NLP**), along with the traditional statistical techniques: simple mean (SM), MSMES, MSMES-P, BMA and BMA-P, in both the (a) test and (b) training sets for the 317 locations. The BESTFCST box is the MAE of the best ensemble member. Fig. 3 is similar to Fig. 2 for the 72-h forecast, and includes GGP's decoupled approach (GGP_2) for the four grammars (**L**, **LP**, **NL** and **NLP**).

Figs. 2a and 3a reveal the superiority of three GP versions (GE_1, GE_2 and GGP_2) over the statistical techniques: simple mean, MSMES and MSMES-P, since the GP boxes do not overlap the last three ones, with medians roughly 27–57% lower than simple mean (6.84 mm (24-h); 7.14 mm (72-h)), MSMES (6.65 mm (24-h); 6.96 mm (72-h)) and MSMES-P (6.63 mm (24-h); 6.92 mm (72-h)). On the other hand, GE_1, GE_2 and GGP_2 are equivalent to each other and to BMA and BMA-P.

Typically the forecast errors get worse as the forecast range increases. Indeed, the medians for the 72-h forecast were 2–5% higher than those for the 24-h forecast (see Figs. 2a and 3a).

Although the BESTFCST, GE and GGP boxes partially overlap, the BESTFCST median (6.17 mm (24-h); 6.44 mm (72-h)) falls above the upper quartile of GEs and GGPs, whose medians fall below the lower quartile of BESTFCST. From the statistical viewpoint, the DEF of 24-h accumulated precipitation via evolutionary algorithms are slightly better than the best individual forecast of the ensemble, reaching an improvement of 34–42%. Such a statement can also be applied to BMA and BMA-P, but nothing can be stated about the other statistical techniques. Even though the ensemble postprocessing via simple mean, MSMES and MSMES-P have not performed better than BESTFCST, there is an advantage to using them since in practice one does not know which is the best ensemble member.

By following the same reasoning, the six grammars (**L**, **LP**, **LA**, **NL**, **NLP** and **NLA**) show a great similarity among themselves, with equivalent interquartile ranges, means, medians and whisker lengths, suggesting that (i) the phenomenon is not significantly non-linear, (ii) the influence of weather patterns over dynamical NWP models is negligible and/or (iii) none of the meteorological attributes play an

[2] http://www.epochx.org/

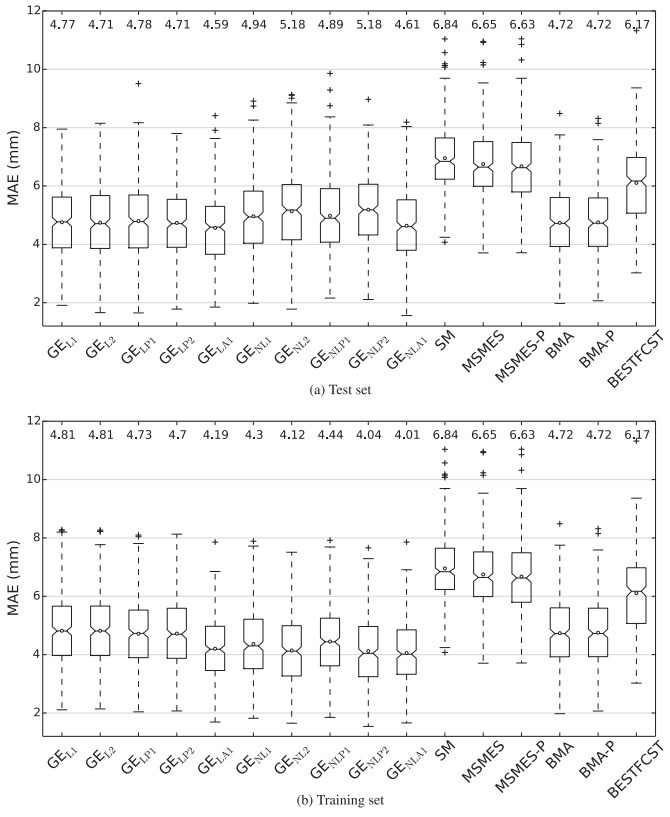

Fig. 2. Box plots of MAE (mm) of the ensemble 24-h QPF achieved through GE_1 for the six grammars (**L**, **LP**, **LA**, **NL**, **NLP** and **NLA**) and GE_2 for the four grammars (**L**, **LP**, **NL** and **NLP**), along with the statistical techniques: simple mean (SM), MSMES, MSMES-P, BMA and BMA-P, in the (a) test and (b) training sets for the 317 locations. The BESTFCST box is the MAE of the best ensemble member. Open circles inside the box represent the sample mean. Median values are displayed at the top of the graph.

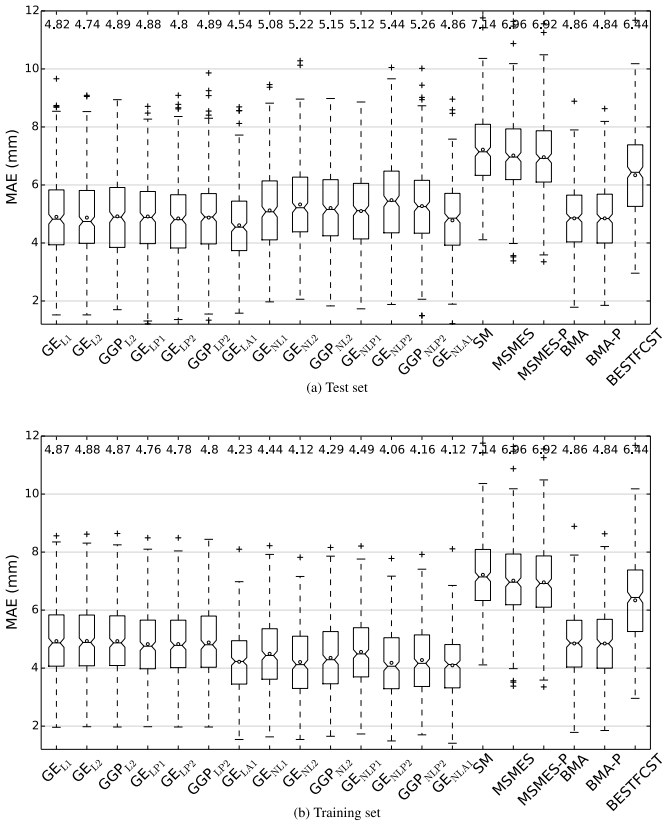

Fig. 3. Box plots of MAE (mm) of the ensemble 72-h QPF achieved through GE_1 for the six grammars (**L**, **LP**, **LA**, **NL**, **NLP** and **NLA**), GE_2 and GGP_2 for the four grammars (**L**, **LP**, **NL** and **NLP**), along with the statistical techniques: simple mean (SM), MSMES, MSMES-P, BMA and BMA-P, in the (a) test and (b) training sets for the 317 locations. The BESTFCST box is the MAE of the best ensemble member. Open circles inside the box represent the sample mean. Median values are displayed at the top of the graph.

important role in bringing major improvements to ensemble postprocessing. The first hypothesis is supported by the fact that the **NL** and **NLP** medians were 2–15% higher than those of the **L** and **LP** grammars in the test set (see Figs. 2a and 3a). The second hypothesis assumes that the four predefined weather patterns represent the clearly distinguishable major weather patterns of Brazil.

For a context-free grammar (Chomsky, 1956), a generally accepted measure of grammatical complexity is based on the number of operators,[3] operands[4] and production rules in the grammar. For any given grammar, its complexity increases as operators, operands and production rules are added. Consequently, the search space size and the difficulty of the problem also increase. Thus, we can conclude that the **L** grammar is less complex than both **LP** and **NL** grammars, which are less complex than the **NLP** grammar, which in turn is less complex than the **NLA** grammar. It still remains that **LA** is more complex than the **L** and **LP** grammars, and less complex than the **NLA** grammar. However, the same computational effort was employed for all the grammars. Thus, an alternative hypothesis attributes the equivalence between the grammars to the underestimation of the computational effort required for an adequate exploitation of their respective search spaces.

According to Figs. 2b and 3b, the boxes of the non-linear grammars (**NL** and **NLP**) shifted toward lower values of the MAE when compared to linear grammars (**L** and **LP**), with medians 6–18% lower than those of the linear grammars. Additionally, the **NL** and **NLP** medians in the training set are reduced 10–34% relative to the medians in the test set, in contrast to just 1–2% for the **L** and **LP** grammars. The **NLA** medians in the training set are reduced 15–18% relative to the medians in the test set, and 7–10% for the **LA** grammar. Based on these results, another alternative hypothesis in order to explain the equivalence between linear and non-linear grammars is that the complexity of the **NL**, **NLP**, **NLA** and **LA** grammars may have led to overfitting. Note that regardless of whether or not overfitting has occurred, it does not compromise our interpretation of Figs. 2a and 3a once the test MAEs are satisfactory. Furthermore, some steps can be taken in order to reduce overfitting and further improve the performance, such as: (i) increasing dataset size, (ii) minimizing noise in the dataset, (iii) and adjusting the GP parameters. It is worth remembering that the selection strategy of the ensemble members was based on the MSMES algorithm. Hence, MSMES and, indirectly, the other linear methods, such as simple mean, MSMES-P, BMA, BMA-P and linear grammar-based evolutionary algorithms, have a clear advantage over non-linear grammar-based evolutionary algorithms.

Although the results for the six grammars are almost indistinguishable, the **LA** grammar-based GE_1 box had a median 3–13% lower than the other GE and GGP boxes for the 24-h QPF, and 4–20% for the 72-h forecast. Notice that the GE_{LA1} and GE_{NLA1} medians are nearly equal for the 24-h forecast, therefore the last one was not taken into account for calculating the above range of percentage improvement. With respect to BMA and BMA-P, the GE_{LA1} median showed an improvement of 3% and 7% for the 24- and 72-h forecasts, respectively. For the 72-h forecast, the GE_{LA1} median is significantly different from BMA and BMA-P medians at the 5% level, since their notches do not overlap (Chambers et al., 1983): the upper notch of GE_{LA1} (4.54 mm) falls below the lower notches of BMA (4.72 mm) and BMA-P (4.69 mm).

[3] Examples of operators include mathematical, logical, relational and conditional operators as well as iterative loops.

[4] Examples of operands may be input attributes, numerical constants and functions with no arguments, such as the function `rand()`, which returns random numbers.

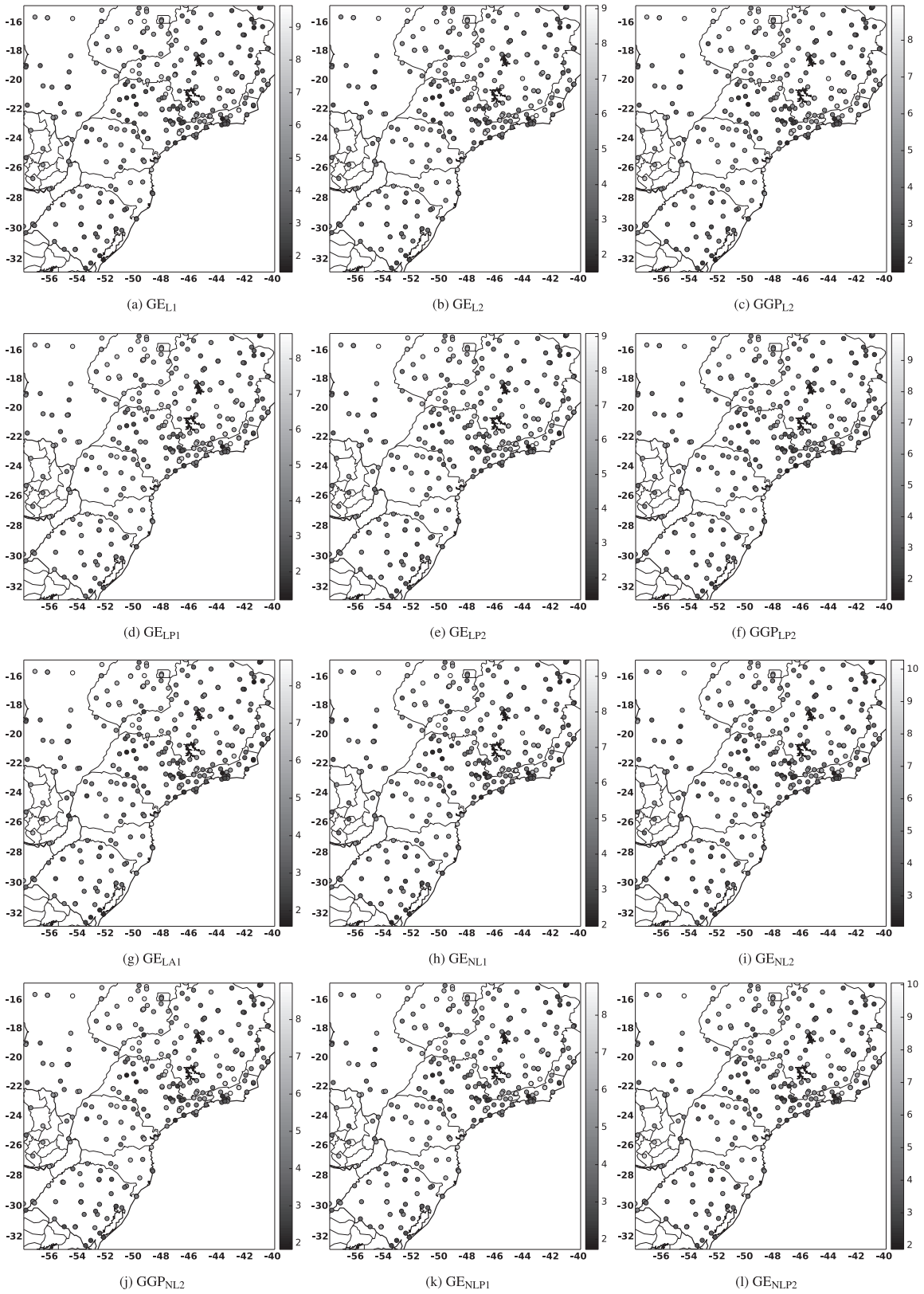

Fig. 4. Geographic distribution maps of the MAE (mm) value for each of the 317 locations for the ensemble 72-h QPF obtained from GE_1 for the six grammars (**L**, **LP**, **LA**, **NL**, **NLP** and **NLA**), GE_2 and GGP_2 for the four grammars (**L**, **LP**, **NL** and **NLP**), along with simple mean (SM), MSMES, MSMES-P, BMA, BMA-P and BESTFCST, in the test set.

In addition to the analysis for the entire spatial domain, the continental region was divided into sixteen homogeneous rainfall sub-regions by applying K-means unsupervised clustering method (MacQueen, 1967) to time series of daily quintiles from October to February on 91 × 91 = 8281 grid points. The daily quintiles of rainfall amount were calculated from a 3-day window centered on each calendar day in the base period 1998–2013. The measure of similarity was based on Pearson correlation coefficient. Each of the 317 locations was assigned to one of the sixteen sub-regions according to its geographical position. Box plots were constructed separately for each

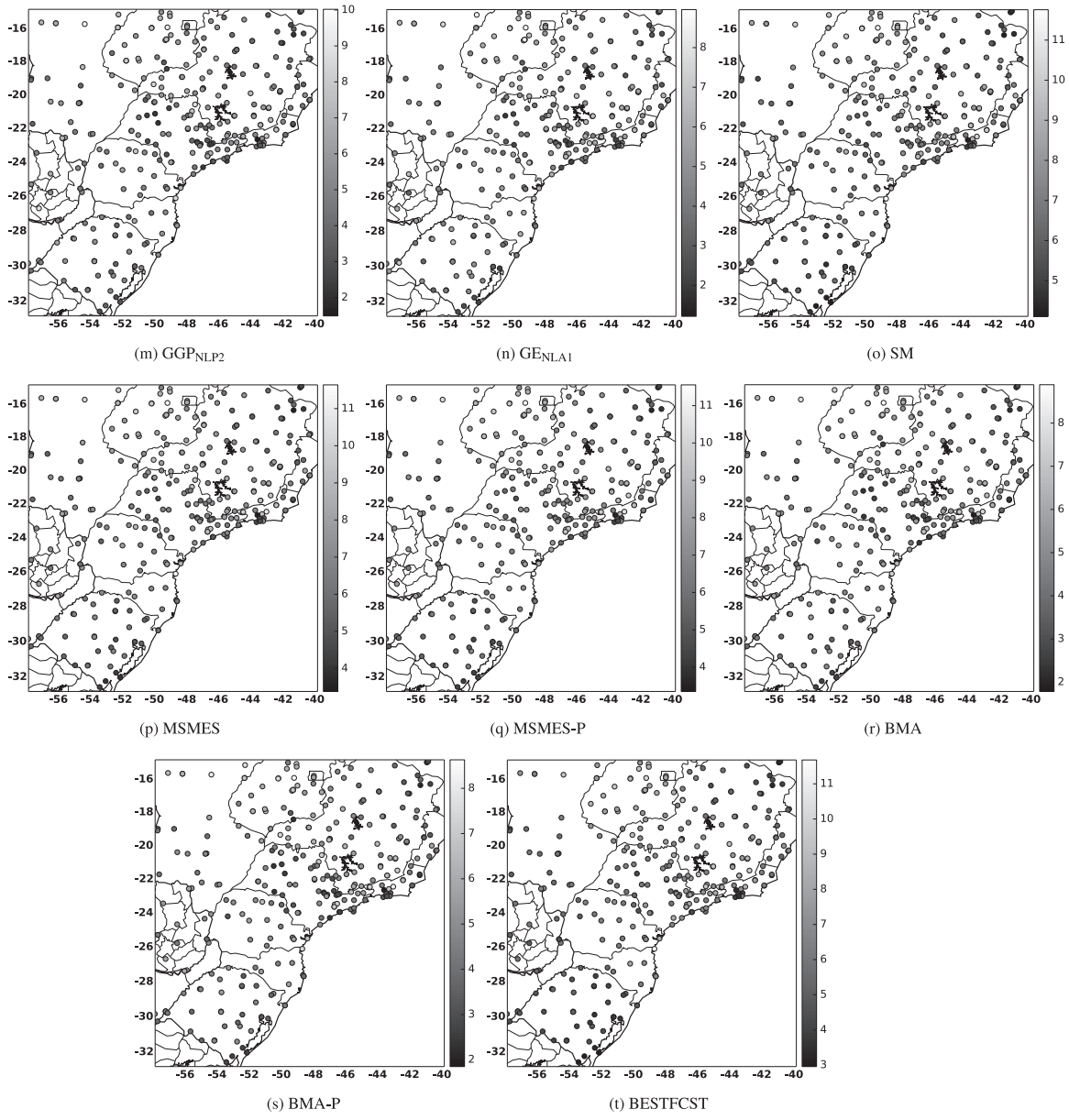

Fig. 4. (*continued*)

sub-region of the domain, including only the MAEs relative to the locations belonging to the sub-region in question. From that, it was observed that in general the behavior of the sixteen homogeneous rainfall sub-regions is similar to that obtained for the entire spatial domain (not shown).

3.3.2. Spatial analysis

Fig. 4 shows the geographic distribution maps of the MAE value for each of the 317 locations for the ensemble 72-h QPF obtained from GE_1 for the six grammars (**L**, **LP**, **LA**, **NL**, **NLP** and **NLA**), GE_2 and GGP_2 for the four grammars (**L**, **LP**, **NL** and **NLP**), and also simple mean (SM), MSMES, MSMES-P, BMA, BMA-P and BESTFCST, for the test set. There are strong agreements among all maps: the largest MAE values are concentrated over parts of São Paulo (SP), southern Rio de Janeiro (RJ) and Minas Gerais (MG), and Goiás (GO) states, while the lowest ones are over Rio Grande do Sul (RS), central and southern São Paulo (SP), and central and northern Minas Gerais (MG) states, including Espírito Santo (ES) and northern Rio de Janeiro (RJ) states. The location of the Brazilian states is shown in Fig. 1.

Calculating the difference between the MAE of the ensemble 24- and 72-h QPF obtained from GE_{LA1} and BMA in the test set (not shown), we found that roughly 65% of the differences are negative, i.e. GE_{LA1} had the MAE lower than BMA's. The difference between the MAE of the ensemble 24- and 72-h QPF obtained from GE_{LA1} and GE_{NLA1} in the test set shows that roughly 60% of the differences are negative, i.e. GE_{LA1} forecasts were more accurate than GE_{NLA1} ones (not shown).

Based on these results, one can conclude that the best technique and/or grammar varies according to the location and the forecast range. Therefore, the techniques are not mutually exclusive, but complementary, since no technique outperformed the others in all locations under all circumstances.

3.3.3. Symbolic solution

As mentioned before, one of the major advantages of the GP algorithms is their capacity of evolving human-interpretable solutions, thus allowing the direct extraction of knowledge from them. In this section, we present the summary outputs of a sensitivity analysis in a table form, as well as a brief analysis of a selected symbolic solution.

A sensitivity analysis allows one to assess the impact that changes in

Table 5
Normalized sensitivity median (%) and the relative frequency (%) of 35 attributes in the 3 170 possible solutions given by GE_{LA1} and GE_{NLA1} to the ensemble 24- and 72-h forecast problem. In parentheses is the rank of the values of the respective column in descending order. The attributes are sorted according to the consensus ranking achieved with the Borda count method applied separately to each grammar (**LA** and **NLA**).

GE_{LA1}					GE_{NLA1}				
Attribute	**Normalized Median (%)**		**Relative Frequency (%)**		**Attribute**	**Normalized Median (%)**		**Relative Frequency (%)**	
	24-h	**72-h**	**24-h**	**72-h**		**24-h**	**72-h**	**24-h**	**72-h**
BMA	18.27(1)	15.78(1)	86.47(1)	86.72(1)	BMA	14.07(1)	11.67(1)	84.64(1)	84.48(1)
rain	9.85(2)	10.30(2)	21.58(5)	21.80(5)	rain	5.33(2)	7.02(2)	29.31(4)	34.07(4)
pattern_change	5.63(4)	6.42(3)	22.18(4)	23.19(4)	pattern_change	4.14(5)	4.56(4)	40.03(3)	39.91(3)
P	4.60(7)	4.82(7)	28.49(2)	30.06(2)	P	3.07(10)	3.44(7)	42.02(2)	41.48(2)
M(min)	3.36(8)	3.86(8)	25.71(3)	27.19(3)	M(min)	3.43(6)	4.23(5)	21.48(5)	19.81(8)
SWEAT	5.60(5)	5.59(5)	16.62(6)	20.25(6)	SWEAT	3.09(9)	3.83(6)	21.29(6)	23.44(5)
TT	6.02(3)	6.33(4)	8.77(15)	7.57(17)	TT	4.79(3)	4.99(3)	10.22(21)	10.03(22)
O(lag3-)	2.30(11)	2.18(15)	11.01(10)	9.59(14)	SFENM	2.61(14)	2.47(13)	15.39(12)	13.72(15)
O(lag1-)	1.98(14)	2.35(12)	7.60(18)	7.85(15)	ACOPL	2.35(19)	2.24(19)	19.31(7)	16.59(10)
M(mean)	2.38(10)	2.20(14)	6.62(22)	5.30(24)	O(lag1+)	1.99(23)	2.08(25)	19.12(8)	21.99(6)
SFENM	2.27(13)	1.54(20)	7.38(19)	7.38(19)	ETAcr	2.79(11)	2.65(9)	9.27(24)	9.59(23)
K	5.13(6)	5.00(6)	2.62(31)	2.81(30)	RAMSC	3.34(7)	2.25(17)	1.58(35)	17.07(9)
ACOPL	1.25(27)	1.06(27)	13.28(9)	11.77(11)	M(mean)	2.23(22)	2.31(15)	12.40(18)	12.87(17)
O(lag3+)	1.24(28)	1.40(23)	10.28(12)	10.76(12)	GP213	2.44(16)	2.03(27)	16.75(10)	9.37(25)
O(lag1+)	0.91(32)	0.92(29)	14.76(8)	16.09(8)	ETAm4	3.09(8)	2.47(14)	5.96(29)	6.40(29)
RAMSC	2.41(9)	1.33(24)	1.14(35)	11.86(10)	M(std)	1.83(26)	2.23(20)	14.51(15)	12.56(19)
ETAm1	1.54(22)	2.50(11)	3.97(27)	7.07(20)	K	4.52(4)	2.93(8)	2.56(34)	3.38(35)
M(std)	1.63(20)	1.98(16)	6.53(23)	6.56(21)	O(1day)	1.59(31)	1.60(34)	19.09(9)	20.16(7)
O(lag2-)	1.07(30)	1.66(18)	8.74(16)	7.38(18)	T299x	2.79(12)	2.62(10)	8.99(25)	3.75(34)
O(1day)	0.62(34)	0.55(35)	15.14(7)	18.58(7)	ETAm2	2.38(18)	2.62(11)	6.72(26)	7.95(27)
ETAcr	1.34(23)	0.87(30)	7.85(17)	7.60(16)	O(lag3-)	1.72(28)	1.83(28)	13.75(16)	15.55(11)
O(lag2+)	0.62(35)	0.83(32)	10.32(11)	12.05(9)	O(lag3+)	1.93(24)	1.75(31)	14.83(14)	13.44(16)
ETAm2	1.31(25)	1.81(17)	4.67(24)	5.39(23)	O(2days)	1.30(34)	1.76(29)	15.08(13)	14.73(13)
GP213	1.86(16)	0.74(33)	8.99(14)	4.64(27)	O(mean)	1.59(30)	2.29(16)	9.94(22)	10.22(21)
O(P)	1.98(15)	2.28(13)	2.24(33)	2.78(31)	O(lag2+)	1.38(33)	1.71(33)	15.93(11)	15.17(12)
O(2days)	0.72(33)	0.69(34)	9.15(13)	10.13(13)	O(lag1-)	1.78(27)	1.75(30)	12.21(19)	13.82(14)
SFAVN	2.28(12)	1.58(19)	3.97(28)	2.02(34)	SFAVN	2.38(17)	2.55(12)	5.99(28)	4.07(33)
CPTEC	1.84(17)	3.62(9)	2.18(34)	1.51(35)	M(max)	1.50(32)	2.24(18)	11.55(20)	11.10(20)
M(max)	1.60(21)	0.86(31)	6.66(21)	5.43(22)	O(P)	2.26(21)	2.10(24)	9.40(23)	9.40(24)
ETA20	1.84(18)	1.11(26)	4.23(26)	3.75(28)	ETA20	2.77(13)	2.08(26)	6.40(27)	6.44(28)
ETAm4	1.33(24)	1.22(25)	4.38(25)	5.02(25)	ETAm1	2.29(20)	2.20(21)	5.24(31)	9.27(26)
T299x	1.67(19)	0.97(28)	7.10(20)	2.11(33)	ETAm3	2.53(15)	2.11(23)	3.53(33)	5.77(30)
O(mean)	1.29(26)	1.46(21)	3.50(29)	3.53(29)	O(lag2-)	1.22(35)	1.44(35)	13.28(17)	12.59(18)
RPSAS	1.07(31)	2.75(10)	2.62(32)	2.62(32)	CPTEC	1.64(29)	2.18(22)	5.33(30)	5.17(31)
ETAm3	1.20(29)	1.46(22)	2.93(30)	4.67(26)	RPSAS	1.85(25)	1.75(32)	5.24(32)	4.89(32)

input attributes will have on the daily rainfall amount predicted by GP models. The sensitivity analysis was performed as follows. Let the attributes related to daily rainfall amount be approximated by a gamma distribution with shape and scale parameters equal to, respectively, 0.5 and 5, and the three atmospheric instability indices (**K**, **TT** and **SWEAT**) uniformly distributed over the interval [0,30], [0,60] and [0,600], respectively, which is the range of values typically observed. 100 randomly sampled input scenarios—i.e. all input variables are simultaneously sampled—were generated for each of the ten solutions at each location (317 in total). For each scenario, one attribute at a time is randomly changed, and the absolute difference between the results generated by the GP model before and after the change was recorded. This procedure is repeated 30 times for each of the 35 attributes (21 meteorological variables and 14 NWP models), except for the discrete-valued attributes **pattern_change**, **rain** and **P**, where all possible input values were assessed. The sensitivity analysis were carried out for each one of the two grammars (**LA** and **NLA**) for each forecast range (24- and 72-h).

Table 5 shows the normalized sensitivity median and the relative frequency of the 35 attributes considering the 317 locations × 10 runs = 3 170 possible solutions given by GE_{LA1} and GE_{NLA1} to the ensemble 24- and 72-h forecast problem. The relative frequency of a certain attribute was defined based on the following criterion: if the mean from a sample of 100 medians, which in turn were calculated from a sample of 30 absolute differences, is non-zero, then the solution statistically includes in its structure the attribute in question, otherwise it does not include it.[5] Following the same criterion, only the means with non-zero values were taken into account for calculating the normalized sensitivity median of an attribute. In parentheses is the rank of the values of the respective column in descending order. In Table 5, the input attributes are sorted according to the consensus ranking achieved with the Borda count method (Borda, 1784) applied separately to each grammar (**LA** and **NLA**), with four different rankings for each one.

According to Table 5, roughly 85% of the programs evolved by the GE_{LA1} algorithm included in their structures the **BMA** attribute, as expected since it is a good linear combination of models. However, the inclusion of the **BMA** in the linear grammar (**LA**) might have impaired the exploration of the search space once it provides a very good approximation (local optimum) which leads to strong suboptimal attraction basins. Similar results were obtained for the two GE_{NLA1} experiments.

Using the Borda count method, a consensus ranking of the 35 attributes is obtained by aggregating information from the eight different rankings provided by Table 5. In the final overall ranking, **BMA** is ranked first; **rain** is ranked second; **pattern_change** is ranked third, followed by **P** and **M(min)**. Therefore, these attributes seem to play an important role in the prediction of daily rainfall amount, being present in over 20% of the final solutions. Although the

[5] This criterion ensures that possibly retained introns—parts of the solution that are noneffective, i.e. do not affect program behavior—are identified and then ignored in order to prevent skewing the analysis.

two atmospheric instability index **SWEAT** and **TT** have been little utilized by the programs with a frequency of usage less than 20%, their normalized sensitivity medians were regularly among the five highest values.

The size of the programs evolved by the GE_{LA1} and GE_{NLA1} algorithms for the short-range ensemble QPF problem varied from 4 to 141 primitives,[6] being 37 primitives the mean size. As an illustration, a program corresponding to ensemble 24-h QPF for Franca-SP from the GE_{NLA1}—called $\mathcal{P}_1$—is shown in Fig. 5. The program $\mathcal{P}_1$ is an example of those whose size is closest to the most common size, with 28 primitives and MAE of 6.98 mm and 5.73 mm in the training and test sets, respectively, and includes in its structure three of the five most important attributes. $\mathcal{P}_1$ uncovered a relationship between 7 of 21 meteorological attributes available; they are: **rain**, **BMA**, **M(max)**, **O(P)**, **O(lag2+)**, **O(lag1-)** and $\mathbf{P_1}$, besides the QPF provided by three (M_1, M_2 and M_3) of the four ensemble members (M_1 = ACOPL, M_2 = ETAm4, M_3 = GP213 and M_4 = SFENM), totaling 10 input attributes. The $\mathcal{P}_1$ program (MAE = 5.73 mm) gives more accurate QPFs for Franca-SP than do the simple mean (MAE = 9.68 mm), MSMES (MAE = 9.41 mm), MSMES-P (MAE = 9.23 mm), BMA (MAE = 7.25 mm) and BMA-P (MAE = 7.24 mm). According to the $\mathcal{P}_1$ program, the **rain** attribute divides the days into two classes: **M(mean)** below or equal to 1 mm, and above 1 mm. For the first class, the 24-h accumulated precipitation was estimated by the **BMA**. The second one has a more complex structure, including eight numerical input attributes, which describes three different paths to be traced, i.e. three different algebraic expressions to predict daily rainfall amount.

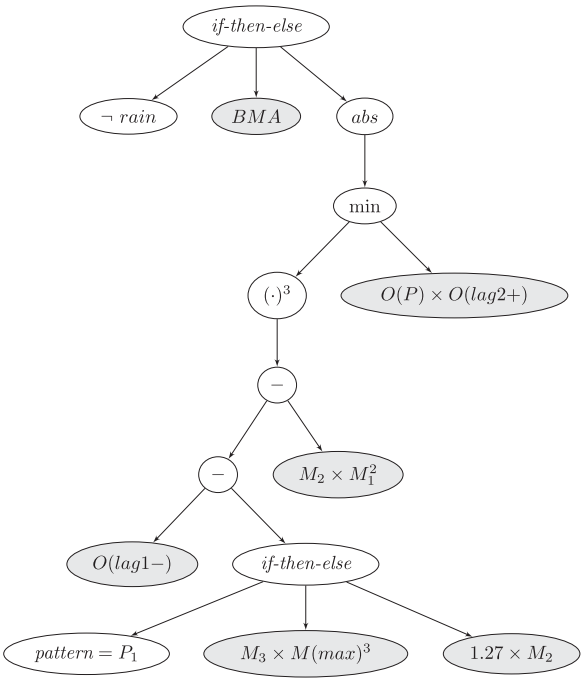

Fig. 5. Expression-tree representation of the program $\mathcal{P}_1$ corresponding to ensemble 24-h QPF for Franca-SP from the GE_{NLA1}, with 28 primitives and MAE of 6.98 mm and 5.73 mm in the training and test sets, respectively. The filled-in gray ellipses are external nodes, also known as leaf nodes. The expression tree is read from left to right, starting at the top and working downward.

4. Conclusions and future work

GP algorithms were explored in order to provide a more accurate and reliable deterministic ensemble forecasts of 24-h accumulated precipitation. Three GP versions and six ensemble forecast grammars were applied to 24- and 72-h forecasts at 317 locations in southern, southeastern and central Brazil during the rainy period from October to February of 2008–2013.

The GP deterministic ensemble forecasts provide substantial improvements in accuracy of rainfall amount forecasting relative to two traditional statistical techniques, with errors 27–57% lower than simple ensemble mean, MSMES and MSMES-P, and are also superior to the best individual forecasts in 34–42%. On the other hand, the three GP versions are equivalent to each other and to BMA and BMA-P. However, even though no formal statistical test has been performed, BMA had mean absolute errors higher than that of GE_{LA1} in 65% of the 317 locations. Furthermore, the grammar-based approaches have some valuable advantages over BMA. One of the benefits of grammar-based GP is its ability to evolve expressions of arbitrary complexity, as opposed to the bounded complexity assumed by BMA. Another potential benefit is the possibility of incorporating domain knowledge into the grammar by biasing the final programs' form and/or predictors. Unlike BMA, the method is not restricted solely to NWP models as inputs. In addition, GP offers human-interpretable solutions, i.e. it reveals the internal structures of all the created models. The white-box characteristic of GP gives an insight into the relationship between input and output data, which is a significant advantage of GP over BMA. In contrast to BMA, GP does not require prior knowledge about the statistical distribution of the data, nor does have shortcomings when handling non-linear problems. Moreover, it is worth mentioning that GE and GGP are conceptually simple techniques, robust, potentially non-linear and easily parallelizable.

The six grammars have shown a great similarity among them performance-wise, suggesting that (i) the phenomenon is not significantly non-linear, (ii) the influence of weather patterns over numerical weather prediction models is negligible, (iii) none of the meteorological attributes play an important role in bringing major improvements to ensemble postprocessing, (iv) the computational effort required for an adequate exploration of their search spaces was underestimated, and/or (v) the complexity of the non-linear grammars (**NL** and **NLP**) may have led to overfitting. The first three hypotheses contradict the intuitive expectations of meteorologists, and thus an investigation into the last two hypotheses as well as about other ways to extract the major weather patterns over Brazil would be interesting as a follow-up work.

Although the results for the six grammars are almost indistinguishable, our most complex linear grammar (**LA**) turned out to be the best overall proposal. For the GE_{LA1} experiments, the two input attributes most often utilized by the programs are: BMA deterministic ensemble forecast of 24-h accumulated precipitation, present in over 85% of them, and weather patterns over Brazil, with about 30% of relative frequency.

In general, the experiments showed the potential of the GP approach and suggest that further research on the improvement of the technique is a promising line of research. This technique is applicable to a wide variety of forecast problems and is extensible to probabilistic forecasting, which in some situations can provide greater utility than deterministic forecasting. Another useful direction would be to add other information as input attributes that could improve GP performance.

Acknowledgments

The authors would like to thank the support provided by CNPq (grants 140680/2010-1, 486103/2012-9, 310778/2013-1 and 502836/2014-8), FAPERJ (grants E26/100.388/2012) and the project IBM/LNCC (B1258534).

References

Augusto, D.A., Barbosa, H.J.C., Barreto, A.M.S., Bernardino, H.S., 2011. Evolving numerical constants in grammatical evolution with the ephemeral constant method.

[6] The allowed operators and operands together define the primitives that are available to the evolutionary process.

In: Antunes, L., Pinto, H.S. editors. Progress in Artificial Intelligence; vol. 7026 of Lecture Notes in Computer Science. Springer Berlin Heidelberg. pp. 110–124.
Bakhshaii, A., Stull, R., 2009. Deterministic ensemble forecasts using gene-expression programming. Weather Forecast. 24, 1431–1451.
Borda, J.C., 1784. Mémoire sur les élections au scrutin. Histoire de l'Académie Royale des Sciences.
Carvalho, L.M., Jones, C., Liebmann, B., 2004. The south atlantic convergence zone: persistence, intensity, form, extreme precipitation and relationships with intraseasonal activity. J. Clim. 17, 88–108.
Chambers, J., Cleveland, W., Kleiner, B., Tukey, P., 1983. Graphical Methods for Data Analysis. Chapman & Hall statistics series; Wadsworth International Group.
Chomsky, N., 1956. Three models for the description of language. IRE Trans. Inf. Theory 2, 113–124.
Cramer, N.L., 1985. A representation for the adaptive generation of simple sequential programs. In: Proceedings of the 1st International Conference on Genetic Algorithms. Hillsdale, NJ, USA: L. Erlbaum Associates Inc. pp. 183–187.
Dufek, A.S., Augusto, D.A., Silva Dias, P.L., Barbosa, H.J.C., 2013. Evaluating the feasibility of grammar-based GP in combining meteorological forecast models. IEEE Congress on Evolutionary Computation.
Eckel, F.A., Mass, C., 2005. Aspects of effective mesoscale, short-range ensemble forecasting. Weather Forecast. 20, 328–350.
Eiben, A.E., Smith, J.E., 2003. Introduction to Evolutionary Computing. SpringerVerlag.
Espinosa, A.M., 2011. Ensemble forecast of rainfall amount in the Rio Grande basin in southeast Brazil during summer 2007/2008. Ph.D. in atmospheric science. University of São Paulo.
Ferreira, C., 2001. Gene expression programming: a new adaptive algorithm for solving problems. Complex Syst. 13 (2), 87–129.
Glahn, H.R., Lowry, D.A., 1972. The use of model output statistics (MOS) in objective weather forecasting. J. Appl. Meteorol. 11, 1203–1211.
Greybush, S.J., Haupt, S.E., Young, G.S., 2008. The regime dependence of optimally weighted ensemble model consensus forecasts of surface temperature. Weather Forecast. 23, 1146–1161.
Koza, J.R., 1992. Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge, MA.
MacQueen, J.B., 1967. Some methods for classification and analysis of multivariate observations. vol. 1. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. pp. 281–297.
O'Neill, M., Dempsey, I., Brabazon, A., Ryan, C., 2003. Analysis of a digit concatenation approach to constant creation. In: Proceedings of the European Conference on Genetic Programming. Springer. pp. 173–182.
Omolbani, M.R., Lee, T.S., Amir, A.D. Review of genetic programming in water resource engineering, 4(11), 2010. pp. 5663–5667.
Otero, F.E.B., Castle, T., Johnson, C.G. 2012. EpochX: Genetic programming in Java with statistics and event monitoring. In: Proceedings of the 2012 Genetic and Evolutionary Conference Companion (GECCO 2012). Philadelphia: ACM Press.
Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using bayesian model averaging to calibrate forecast ensembles. Mon. Weather Rev. 133, 1155–1174.
Roebber, P.J., 2010. Seeking consensus: a new approach. Mon. Weather Rev. 138, 4402–4415.
Roebber, P.J., 2015a. Evolving ensembles. Mon. Weather Rev. 143 (2), 471–490.
Roebber, P.J., 2015b. Using evolutionary programs to maximize minimum temperature forecast skill. Mon. Weather Rev. 143 (5), 1506–1516.
Rozante, J.R., Moreira, D.S., de Goncalves, L.G., Vila, D.A., 2010. Combining TRMM and surface observations of precipitation: technique and validation over South America. Weather Forecast. 25 (3), 885–894.
Ryan, C., Collins, J.J., Collins, J., O'Neill, M., 1998. Grammatical evolution: Evolving programs for an arbitrary language. In: Lecture Notes in Computer Science 1391, Proceedings of the First European Workshop on Genetic Programming. Springer-Verlag. pp. 83–95.
Saha, S., Moorthi, S., Pan, H., 2010. The NCEP climate forecast system reanalysis. Bull. Am. Meteorol. Soc. 91, 1015–1057.
Silva Dias, P.L., Moreira, D.S., Dolif, G., 2006. The master super model ensemble system (MSMES). 8 ICSHMO. pp. 1751–1757.
Sloughter, J.M., Raftery, A.E., Gneiting, T., Fraley, C., 2007. Probabilistic quantitative precipitation forecasting using bayesian model averaging. Mon. Weather Rev. 135, 3209–3220.
Velasco, I., Fritsch, J.M., 1987. Mesoscale convective complexes in the Americas. In: Rainfall Fields: Estimation, Analysis and Prediction. American Geophysical Union. pp. 9591–9613.
Whigham, P.A., 1995. Grammatically-based genetic programming. In: Rosca, J.P. editor. Proceedings of the Workshop on Genetic Programming: From Theory to Real-World Applications. Tahoe City, California, USA. pp. 33–41.
Wilks, D.S., 2006. Statistical Methods in the Atmospheric Sciences; vol. 59 of 2nd Ed. International Geophysics Series. Academic Press. pp. 627.