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

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-