Stoch Environ Res Risk Assess
DOI 10.1007/s00477-014-0890-7

ORIGINAL PAPER

Modelling collocation uncertainty of 3D atmospheric profiles

Rosaria Ignaccolo · Maria Franco-Villoria · Alessandro Fassò

Abstract Atmospheric thermodynamic data are gathered by high technology remote instruments such as radiosondes, giving rise to profiles that are usually modelled as functions depending only on height. The radiosonde balloons, however, drift away in the atmosphere resulting in not necessarily vertical but three-dimensional trajectories. To model this kind of functional data, we introduce a "point based" formulation of an heteroskedastic functional regression model that includes a trivariate smooth function and results to be an extension of a previously introduced unidimensional model. Functional coefficients of both the conditional mean and variance are estimated by reformulating the model as a standard generalized additive model and subsequently as a mixed model. This reformulation leads to a double mixed model whose parameters are fitted by using an iterative algorithm that allows to adjust for heteroskedasticity. The proposed modelling approach is applied to describe collocation mismatch when we deal with couples of balloons launched at two different locations. In particular, we model collocation error of atmospheric pressure in terms of meteorological covariates and space and time mismatch. Results show that model fitting is improved once heteroskedasticity is taken into account.

Keywords Functional linear model · Heteroskedasticity · Generalized additive models · Mixed models · Smoothing · Collocation

1 Introduction

Over the last few years the analysis and modelling of functional data has received an increasing interest, motivated by the availability of dense sets of measurements recorded over some domain, such as time, depth or height, in particular in environmental studies. Cuevas (2014) and Horvàth and Kokoszka (2012) and Zhang's (2013) books provide an up-to-date state of the art in functional data analysis and complement the reference books Ramsay and Silverman (2005) and Ferraty and Vieu (2006). Applications of functional data analysis can be found in various scientific areas, including climatological and environmental ones (see e.g. Ruiz-Medina and Espejo 2012; Escabias et al. 2013; Caballero et al. 2013; Ignaccolo 2013). However, to our knowledge, little reference is made to heteroskedasticity in the functional data literature. In Fassò et al. (2013) an unidimensional heteroskedastic regression model is introduced to deal with the assumption of constant variability not always verified. This is an important topic for two reasons: first, mean estimates need to be adjusted for non-constant variability, and second, modelling the variance function itself is of interest to understand which covariates significantly affect the variance. Wang and Akritas (2010) propose a testing procedure for functional data that assesses the significance of nested effects and their interactions taking into account heteroskedasticity in

Work partially supported by FIRB 2012 grant (project no. RBFR12URQJ) provided by the Italian Ministry of Education, Universities and Research.

R. Ignaccolo (✉)
Dipartimento di Economia e Statistica "Cognetti de Martiis", Università degli Studi di Torino, Torino, Italy
e-mail: rosaria.ignaccolo@unito.it

M. Franco-Villoria
Università degli Studi di Torino, Torino, Italy

A. Fassò
Università degli Studi di Bergamo, Bergamo, Italy

Published online: 15 May 2014

the error terms. In the classical context, to handle heteroskedasticity, in Gijbels et al. (2010) and Nott (2006) a further dispersion parameter is incorporated using the double exponential family of distributions in a generalized linear model framework. Instead, in Karlis et al. (2009) it is suggested to model the variance depending on covariates but they limit the relationship to be linear. In this work, we model the variance in a more flexible way allowing for non-linear effects of the covariates, as already proposed in Fassò et al. (2013), motivated by the same case study.

The availability of atmospheric measurements is becoming larger and larger since high technology radiosondes provide atmospheric profiles of essential climate variables (ECVs), like pressure, temperature, water vapour, wind and aerosol (Immler et al. 2010). The uncertainty of such variables is a key factor for assessing the uncertainty of global change estimates given by numerical prediction models (Thorne et al. 2013). An important source of uncertainty is related to the collocation mismatch in space and time among different observations. Suppose we are comparing two radiosondes, which aim to measure the same environmental variables. Collocation makes reference to the placing of the two instruments in exactly the same geographical location at the same time. Since this is hardly the case, we say that two instruments are imperfectly collocated and the difference between the instruments measurements can be defined as the collocation mismatch.

Understanding collocation mismatch is particularly relevant for atmospheric profiles obtained by radiosondes, as the ballons containing the measuring instruments tend to drift uncontrollably from their initial launch position (see Seidel et al. 2011 and Fassò et al. 2013). In particular, collocation mismatch may depend on potential covariates. While in Fassò et al. (2013) the focus is on relative humidity, here we consider the difference in pressure of coupled launches as the response variable.

It is known that for an isothermal and ideal gas the barometric formula ensures that atmospheric pressure depends only on height (Berberan-Santos et al. 2010). If the two available locations are close enough, it seems reasonable to believe that they are subject to a similar climate regime, and hence any difference in pressure between the profiles should be only noise, independently of meteorological covariates. The available locations in the motivating dataset are about 50 km apart. This distance may be enough for the local meteorological and wind conditions to be different at the two locations. If this is the case, the ideal conditions that the barometric formula assumes are no longer valid, and we may expect a significant impact of covariates.

In the motivating case study, measurements are taken using a balloon that drifts away from its original position as it goes higher up into the atmosphere, so that longitude and latitude coordinates at launch do not remain constant. While in Fassò et al. (2013) the profiles were considered to be vertical (i.e. only dependent on height), now the profiles' trajectories are seen as dependent on longitude, latitude and height, and hence as three-dimensional (3D) profiles. Thus we propose to model atmospheric profiles by considering them as functions of a spatial point $p = (x, y, h) \in P \subseteq \mathbb{R}^3$. This "point based" formulation extends the work done in Fassò et al. (2013), but the proposed modelling strategy incorporates potential heteroskedasticity by means of an iterative algorithm (following Ruppert et al. 2003) that was not considered in Fassò et al. (2013). As a result, covariates estimates can be adjusted for non-constant variability and estimation of the functional mean is improved. Simultaneously the conditional variance is explicitly modelled, allowing to identify significant covariates on the collocation second order uncertainty.

The paper is organized as follows. Section 2 introduces the motivating dataset, while Sect. 3 describes the proposed point based model for 3D functional data. In particular, the model reformulation as a GAM/mixed-model is detailed in Sect. 3.1, whereas Sect. 3.2 presents the estimation procedure by an iterative algorithm that permits to handle heteroskedasticity. In Sect. 4 the model is then illustrated on the motivating case study, where the interest is in collocation error of atmospheric pressure, and conclusions follow in Sect. 5.

2 Motivating dataset

The dataset used in this work is the same as in Fassò et al. (2013) where the interested reader can find further details. It consists of radiosounding profiles of ECVs measured at the Howard University research site in Beltsville, Meryland, USA (39.054°, −76.877°, 88 m a.s.l.), which is also a GRUAN site (GCOS Reference Upper-Air Network, see www.gruan.org and Thorne et al. (2013)), and the US National Weather Service operational site at Sterling, Virginia, USA (38.98°, −77.47°, 53 m a.s.l.). These two sites are sufficiently close, 52 km line distance, and represent a similar climate regime. Figure 1 shows trajectories of the two balloons for height values in 100–10000 m as they drift away from their initial launch position; note that one of the profiles in the Sterling site departs considerably from its initial position already at 100 m.

Moreover we can be confident that for height ranging in 100–10000 m there is no instrumental bias. In fact, Beltsville soundings are based on RS92-SGP sondes, manufactured by Vaisala, while Sterling uses Sippican LMS6 sondes. Differences—at the same height—in the sounding of the atmosphere among sensor types were analysed

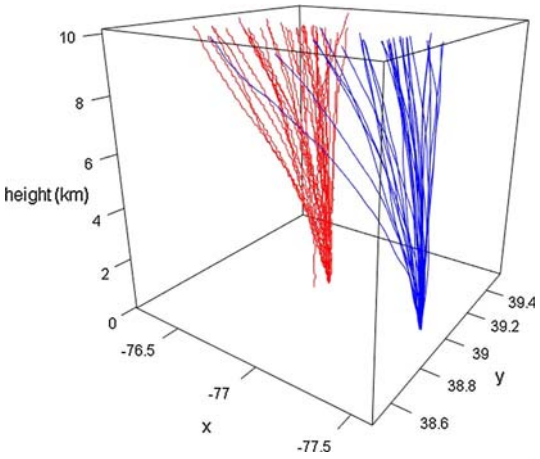

Fig. 1 Trajectories of the two radiosonde balloons launched at Sterling (*blue*) and Beltsville (*red*); x and y denote longitude and latitude in degrees, height ranges from 0.1 to 10 km

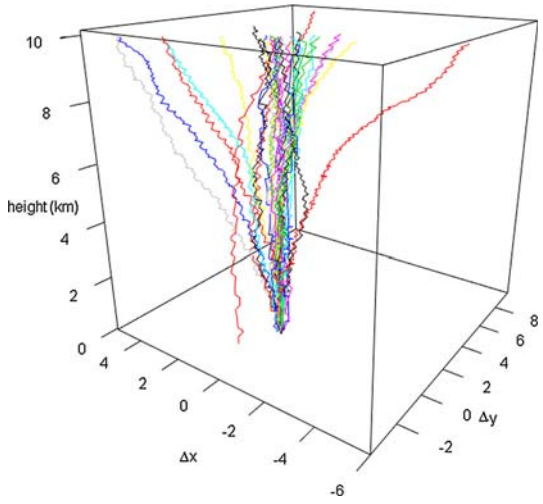

Fig. 2 Differences, for each value of height, in longitude and latitude (Dx and Dy respectively) between the coupled trajectories; height ranges from 0.1 to 10 km

during the last World Meteorological Organization intercomparison of high quality radiosonde systems, as reported in Nash et al. (2010). Both RS92-SGP and LMS6 sondes have been ranked with score 5, that is "Performance ideal for GRUAN", in both cases of pressure higher and lower than 100 hPa.

The Sterling site is considered as the "base" site; in what follows, variables with the superscript "^{0}" (e.g. T^0) refer to the Sterling site, while variables preceeded by "D" make reference to the difference between the two matched profile measurements. A flight from Beltsville was matched to Sterling if launch time was within 3 h. The differences, for each value of height, in longitude and latitude (Dx and Dy respectively) between the coupled trajectories are shown in Fig. 2, that highlights the separation between the two ballons ascending in the atmosphere. We used 32 pairs launched between July 2006 and September 2009; given that the different launches are well spaced in time, it is reasonable to consider the corresponding profiles to be independent.

For each flight, we use data profiles on relative humidity, water vapor mixing ratio, pressure, temperature, measurement calendar time, flight duration, wind vector, distance and coordinates.

3 Modelling 3D atmospheric profiles

Let z denote the measurement of a physical quantity, e.g. an ECV, along a trajectory through the atmosphere. A measurement is gathered at a spatial point $p = (x, y, h) \in P \subseteq \mathbb{R}^3$ and time s, where x, y and $h \geq h_0$ are the measurement longitude, latitude and height, $\mathrm{s} \geq \mathrm{s}_0$ is measurement time, while s_0 and h_0 are launch time and height; moreover the launch place will be denoted as $s = (x_0, y_0, h_0)$. The spatial trajectories can be described as profiles or functions with three-dimensional domain, that is $z_j : \mathbb{R}^3 \to \mathbb{R}$, labelled by launch place and time s_j, s_{0j}, $j = 1, \ldots, n$. According to the functional data analysis approach described e.g. by Ramsay and Silverman (2005), we consider a profile as a single object described by a smooth function $\mathrm{l}(p)$. Motivated by the case study we assume that, for each profile, there is a one-to-one correspondence between p and h, thus we can refer to each point of the trajectory by specifying its height h.

According to standard measurement error decomposition, an observation profile is given by a random function

$$z_j(\cdot) = \mathrm{l}_j(\cdot) + \mathrm{e}_j(\cdot),\ j = 1, \ldots, n, \tag{1}$$

where $\mathrm{l}(\cdot)$ is the "true" profile, assumed to be smooth taking values in a separable Hilbert space, and $\mathrm{e}(\cdot)$ is the zero mean measurement error with variance $\mathrm{r}^2_{\mathrm{e}}(\cdot)$. Figure 3 shows examples of such 3D smoothed functional data for climate variables at the "base" site (Sterling) in our case study, where smoothing in (1) is carried out by penalized cubic B-splines.

In this paper we suppose that measurements, conditionally on a set of forcing factors, are independent random functions and we focus on modelling their conditional mean and variance. In particular, we are interested in comparing two instruments, e.g. radiosondes, launched at two close spatial sites at the same height. The base site trajectories will be described by $p^0 = (x^0, y^0, h)$ whereas the other site by $p = (x, y, h)$, and the associated measurements will be denoted by $z^0 = z(p^0)$ and $z = z(p)$ respectively. Since we are interested in taking differences

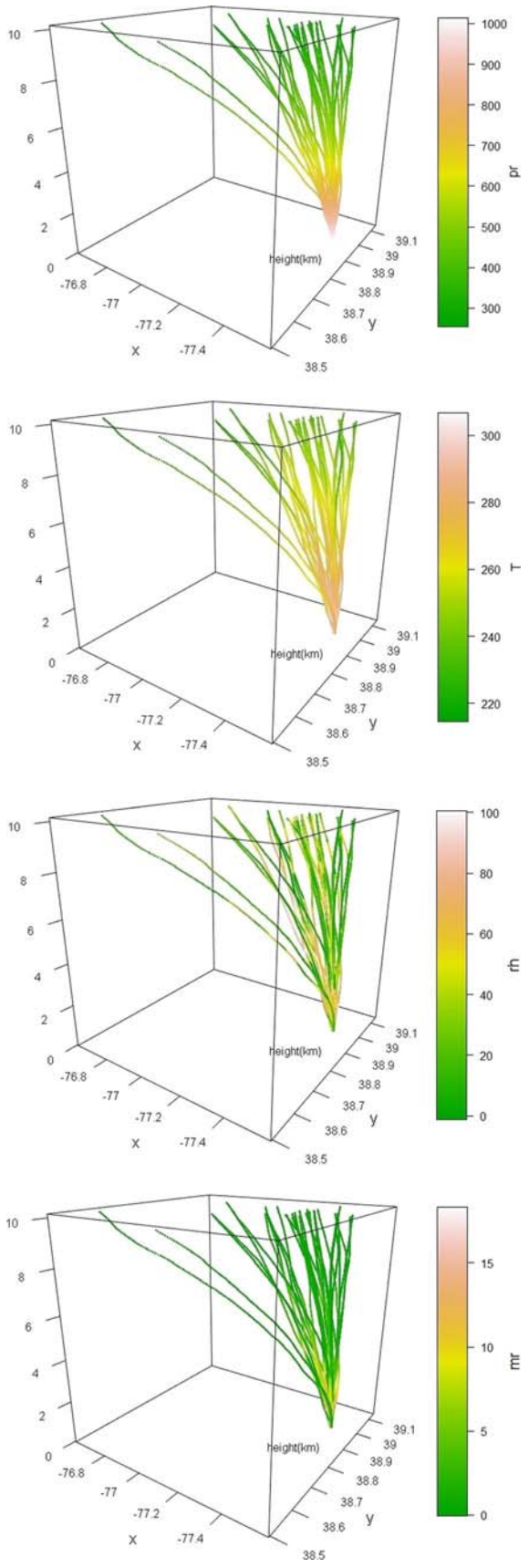

Fig. 3 Atmospheric profiles in 3D for pressure (*pr*), temperature (*T*), relative humidity (*rh*) and water vapour mixing ratio (*mr*). Each *curve* represents a different launch at the base site (Sterling)

at the same height, we will denote $\mathrm{D}p = (\mathrm{D}x, \mathrm{D}y, h)$ where $\mathrm{D}x = x - x^0$ and $\mathrm{D}y = y - y^0$. Then we move from the couple (p^0, p) to the couple $(p^0, \mathrm{D}p)$ and, by taking differences, from (1) we have

$$\mathrm{D}z := z - z^0 = \mathrm{Dl} + \mathrm{De} \tag{2}$$

where $\mathrm{Dl} = \mathrm{l} - \mathrm{l}^0$ is the collocation drift and $\mathrm{De} = \mathrm{e} - \mathrm{e}^0$ is the collocation measurement error and we assume the measurement error at the two sites to be equal so that $Var(\mathrm{De}) = \mathrm{r}^2_{\mathrm{e}} + \mathrm{r}^2_{\mathrm{e}^0} = 2\mathrm{r}^2_{\mathrm{e}}$.

Note that since the paired balloons work independently of each other, the collocated profiles z and z^0 are not observed exactly at the same height h, while l and l^0 are continuous functions whose values are obtained after a smoothing step following (1)—e.g. by means of penalized cubic B-splines—and thus Dl may be easily computed for every height h in a common grid for all profiles.

The potential effect of (functional) environmental factors $c(\cdot)$ on the collocation drift Dl can be investigated by means of a functional trend model given by

$$\mathrm{Dl}(\cdot) = m(\cdot) + \mathrm{b}(\cdot)'c(\cdot) + \mathrm{x}(\cdot) \tag{3}$$

where the argument $(\cdot)$ is the couple $(p^0, \mathrm{D}p)$ for all terms except b and c, so that $m(\cdot)$ is a function of $(p^0, \mathrm{D}p)$ and $\mathrm{b}(\cdot)'c(\cdot) = \mathrm{b}_0 + \sum_{q=1}^{Q} \mathrm{b}_q(\cdot)c_q(\cdot)$ with Q being the number of covariates. Actually, in this work, we consider functional coefficients b only dependent on the height h for the sake of model parsimony, so that we will not have $\mathrm{b}(p^0, \mathrm{D}p)$ but $\mathrm{b}(h)$. In addition, given the one-to-one correspondence between h and p in the profiles, $c(p^0, \mathrm{D}p)$ can be considered as $c(h)$. Given that both p^0 and $\mathrm{D}p$ have the same height h, Model (3) can be seen as a "concurrent" functional linear model with respect to h, since the relationship is established at the same h and we assume that the trend is locally linearly related to c but the global relation is not assumed linear.

Moreover the error $\mathrm{x}(\cdot)$ is assumed to be an heteroskedastic component with conditional variance given by $\mathrm{r}^2_{\mathrm{x}}(\cdot|c) = Var(\mathrm{x}(\cdot)|c)$, which is assumed to depend log-linearly on c and a function of $(p^0, \mathrm{D}p)$ denoted by o. Hence we have

$$\mathrm{r}^2_{\mathrm{x}}(\cdot|c) = \exp\left[o(\cdot) + \mathrm{c}(\cdot)'c(\cdot)\right] \tag{4}$$

where $\mathrm{c}(\cdot)'c(\cdot) = \mathrm{c}_0 + \sum_{q=1}^{Q} \mathrm{c}_q(\cdot)c_q(\cdot)$, Q is the number of covariates and—similarly to the case of the trend model above—c and c are considered depending only on h so that

in the following we will have $\mathrm{c}(h)$ and $c(h)$. This skedastic model describes the uncertainty unaccounted for by the collocation drift.

So Eqs. (3) and (4) define an Heteroskedastic Functional Regression Model (HFRM). In both components of HFRM, we need to specify the first term, $m(\cdot)$ and $o(\cdot)$ respectively. We consider three alternatives (written here only for $m(\cdot)$ to avoid repetition):

(A) $m(p^0, \mathrm{D}p) = m_0(p^0) + m_{\mathrm{D}}(\mathrm{D}p)$,
(B) $m(p^0, \mathrm{D}p) = \mathrm{a}_1(h)\, x^0 + \mathrm{a}_2(h)\, y^0 + m_{\mathrm{D}}(\mathrm{D}p)$,
(C) $m(p^0, \mathrm{D}p) = \mathrm{a}_1(h)\, x^0 + \mathrm{a}_2(h)\, y^0 + \mathrm{a}_3(h)\, \mathrm{D}x + \mathrm{a}_4(h)\, \mathrm{D}y + \mathrm{a}_5(h)$,

where h, as above, denotes the common height of p^0 and $\mathrm{D}p$. The three alternatives offer different ways of incorporating longitude, latitude and their differences (that are related to the distance between two points of paired profiles), as well as height in the model, from a more complex and less parsimonious model to a simpler one. Indeed, case (C) treats longitude and latitude, and their differences as covariates and includes a functional intercept ($\mathrm{a}_5(h)$) resulting in a vertical profiles modelling strategy as in Fassò et al. (2013). In case (A), to be consistent with our point based formulation, we consider trivariate functions—namely m_0 and m_{D}—that take into account a possible interaction among longitude, latitude and height, as well as among the distance in terms of $\mathrm{D}x$ and $\mathrm{D}y$, and h. Finally, case (B) represents a middle alternative that allows to have a simpler model but keeps the interaction among $\mathrm{D}x$, $\mathrm{D}y$ and height h.

3.1 GAM and mixed model representation

In order to estimate the function $m(\cdot)$, and then $o(\cdot)$, as well as the functional coefficients $\mathrm{b}(\cdot)$, and then $\mathrm{c}(\cdot)$, we rewrite the above functional linear models as standard (generalized) additive models with penalized splines (following Guo 2004; Ngo and Wand 2004; Harezlak et al. 2007; Ivanescu et al. 2012) moving to a longitudinal data perspective. In fact, as observed in Ramsay and Silverman (2005) p. 258, the concurrent functional linear model in (3) can be seen as a varying-coefficient model (Hastie and Tibshirani 1993) and this is basically a generalized additive model (GAM) where the smooth components are multiplied by known covariates (see Wood 2006 p. 168).

Let us consider HFRM with $m(\cdot)$ as specified in (A) and let us rewrite Model (3) by means of spline basis representation. Both m_0 and m_{D} can be re-expressed in terms of the tensor product of marginal spline basis functions (Wood 2006; Eilers and Marx 2003). To this goal, we consider a smooth function $m_*(x, y, h)$ that in turn will represent m_0, m_{D}, o_0 and o_{D}. For any argument of m_*, i.e. by taking x, y, h individually, and assuming that we have

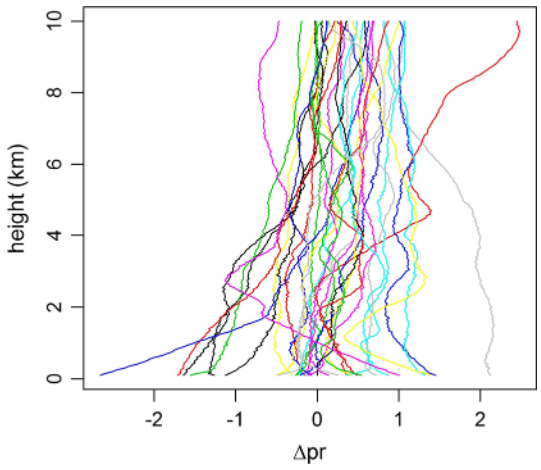

Fig. 4 Collocation mismatch profiles of pressure

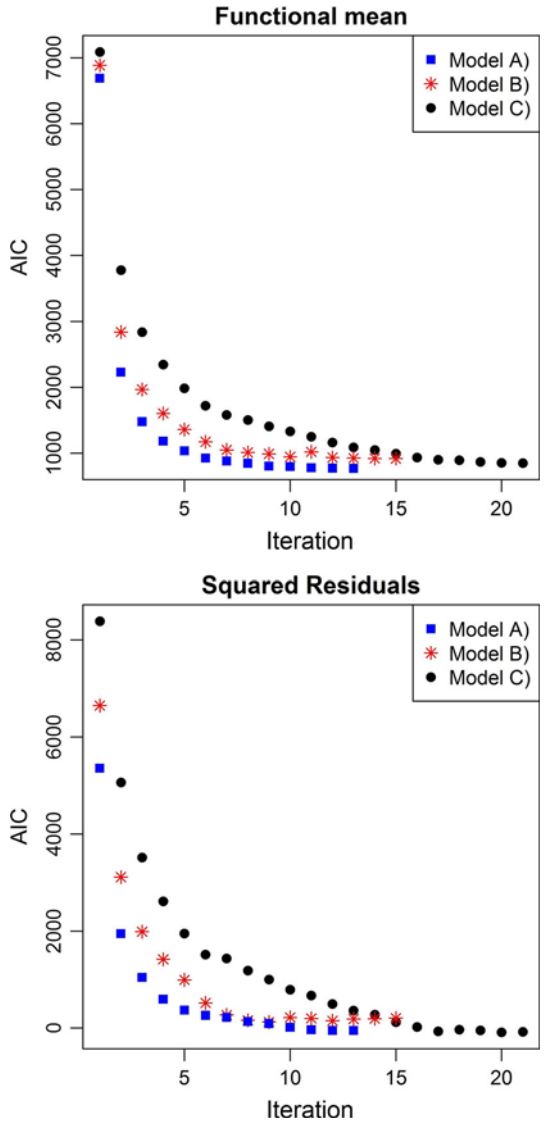

Fig. 5 *AIC* for **f** and *AIC* for **g** by iteration

Table 1 Model summary for the functional mean

Parametric coefficients			
	Estimate	Std. error	p-value
b_0	0.89486	0.04922	\ 2e-16
Smooth terms	**edf**	**F**	**p-value**
$m_0(x^0, y^0, h)$	19.476	156.27	\ 2e-16
$m_D(Dx, Dy, h)$	12.763	76.93	\ 2e-16
s^0	12.459	234.98	\ 2e-16
T^0	13.097	33.57	\ 2e-16
rh^0	12.272	95.92	\ 2e-16
mr^0	12.242	113.00	\ 2e-16
uW^0	11.225	127.56	\ 2e-16
vW^0	12.760	71.21	\ 2e-16
Dt	9.523	132.27	\ 2e-16
Ds_0	10.118	179.35	\ 2e-16
DT	13.422	39.45	\ 2e-16
Drh	2.000	19.52	3.55e-09
Dmr	10.923	10.61	\ 2e-16
DuW	12.921	40.12	\ 2e-16
DvW	13.412	63.19	\ 2e-16

marginal spline basis available $A_{x,k}(x)$, $A_{y,r}(y)$ and $A_{h,v}(h)$, with k_x, k_y, k_h being the number of basis functions for each variable, we can construct the smooth functions $m_1(x), m_2(y)$ and $m_3(h)$ as follows

$$m_1(x) = \sum_{u=1}^{k_x} A_{x,k}(x)\, m_{x,u},$$

$$m_2(y) = \sum_{r=1}^{k_y} A_{y,r}(y)\, m_{y,r},$$

$$m_3(h) = \sum_{v=1}^{k_h} A_{h,v}(h)\, m_{h,v}.$$

By recalling that $p = (x, y, h)$, the trivariate function $m_*(x, y, h)$ can be expressed as

$$m_*(p) = \sum_{u=1}^{k_x}\sum_{r=1}^{k_y}\sum_{v=1}^{k_h} A_{x,u}(x)A_{y,r}(y)A_{h,v}(h)\, m_{xyh,uvr} = \sum_{l=1}^{k} A_{p,l}(p)\, m_{m,l}$$

where $k = k_x \times k_y \times k_h$, $A_{p,l}$ are elements of the $N \times k$ matrix $A_p = A_x \odot A_y \odot A_h$ where $\odot$ denotes the tensor product (for further details see Wood 2006 p. 162) and $N = n \times H$ where H is the number of values of height h when the profiles are discretized (taking their values on a common grid) to move to a longitudinal approach. Moreover we denote $m_m = (m_{m,1}, \ldots, m_{m,k})$ as the vector of spline coefficients (to be estimated).

So for the site s_0 we have $m_0(p^0)$ written as

$$m_0(p^0) = \sum_{l=1}^{k_0} A_{p^0,l}(p^0)\, m_{0,l}$$

where $k_0 = k_{x^0} \times k_{y^0} \times k_h$. Similarly, $m_D(Dp)$ can be written as

$$m_D(Dp) = \sum_{l=1}^{k_D} A_{Dp,l}(Dp)\, m_{D,l}$$

where $k_D = k_{Dx} \times k_{Dy} \times k_h$.

As for the term $b(\cdot)'c(\cdot) = b_0 + \sum_{q=1}^{Q} b_q(\cdot)c_q(\cdot)$ in (3), the functional coefficients $b_q(h)$ are assumed to be expandable as

$$b_q(h) = \sum_{l=1}^{k_q} a_{q,l}(h)n_{q,l}$$

where $a_{q,l}(h)$ are known spline basis functions, while $n_{q,l}$ are the related coefficients (to be estimated). Then we can write

$$b_q(h)c_q(h) = \sum_{l=1}^{k_q} a_{q,l}(h)c_q(h)n_{q,l} = \sum_{l=1}^{k_q} A_{q,l}(h)n_{q,l}$$

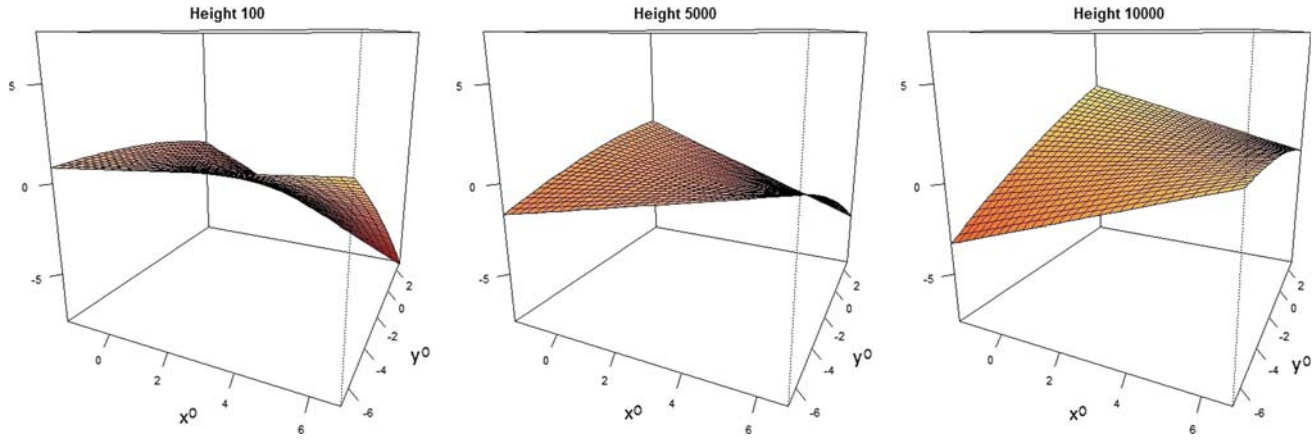

Fig. 6 $\hat{m}_0(x^0, y^0, h)$ for three different values of height

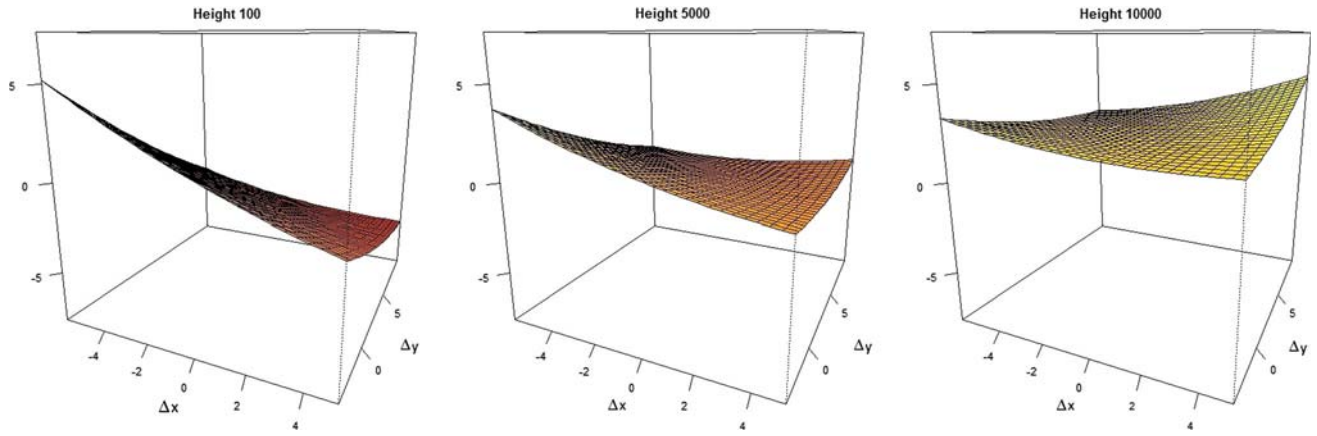

Fig. 7 $\hat{m}_{\mathrm{D}}(\mathrm{D}x, \mathrm{D}y, h)$ for three different values of height

where $A_{q,l}(h) = a_{q,l}(h)c_q(h)$ are known because $c_q(h)$ are "observed" without noise.

Thus the functional linear model (3) can be rewritten as a standard additive model

$$\begin{aligned} \mathrm{D}l\,(p^0, \mathrm{D}p) = & \sum_{l=1}^{K_0} A_{p^0,l}(p^0)\, \mathrm{m}_{0,l} \\ & + \sum_{l=1}^{K_{\mathrm{D}}} A_{\mathrm{D}p,l}(\mathrm{D}p)\, \mathrm{m}_{\mathrm{D},l} + \mathrm{b}_0 \\ & + \sum_{q=1}^{Q} \sum_{l=1}^{K_q} A_{q,l}(h)\mathrm{n}_{q,l} + \mathrm{x}\,(p^0, \mathrm{D}p) \end{aligned} \tag{5}$$

where h is the common height of p^0 and Dp as above. Hence Model (5) corresponds to a GAM with smooth components represented via a regression spline model (Wahba 1990) that is fitted by penalized maximum likelihood estimation to avoid overfitting: a large number of basis functions is chosen and penalties are designed to suppress excessive roughness of the functional parameters. This rewriting and fitting procedure results to be similar to the approach adopted in Harezlak et al. (2007). In practice, the GAM penalized likelihood maximization problem is solved by penalized iteratively reweighted least squares (P-IRLS) to estimate the vector of coefficients

$$\mathrm{n} = \{\mathrm{m}_0, \mathrm{m}_{\mathrm{D}}, \mathrm{b}_0, \mathrm{n}_1, \ldots, \mathrm{n}_Q\}$$

where $\mathrm{m}_0 = \{\mathrm{m}_{0,1}, \ldots, \mathrm{m}_{0,k_0}\}$, $\mathrm{m}_{\mathrm{D}} = \{\mathrm{m}_{\mathrm{D},1}, \ldots, \mathrm{m}_{\mathrm{D},k_{\mathrm{D}}}\}$ and $\mathrm{n}_q = \{\mathrm{n}_{q,1}, \ldots, \mathrm{n}_{q,k_q}\}$, $q = 1, \ldots, Q$. The P-IRLS method assumes the vector of the so-called *smoothing parameters*, multiplying the smooth components' penalties and controlling the trade-off between fidelity to the data and smoothness, to be known (see e.g. Wood 2006, 2011). The estimation of such smoothing parameters can be achieved by minimization of a prediction error estimate, such as the generalized cross validation (GCV) score, or by restricted maximum likelihood estimation (REML) via a mixed effects model representation of a GAM.

The equivalence of smoothing splines and mixed models was highlighted by Speed (1991) while discussing the work of Robinson (1991). Such an equivalence allows for the choice of the smoothing parameters through the estimation of the variance components associated with the random effects in the mixed model representation that are the penalized spline parameters (for details see e.g. Wand 2003; Ruppert et al. 2003; Wood 2004). Reiss and Ogden (2009) show that sometimes REML may be preferable to GCV and Wood (2011) provides a general computationally efficient way of estimating the smoothing parameters that makes REML fast and stable and is implemented in the mgcv package Wood (2012).

Since we deal with collocation uncertainty, we are interested in understanding how variability changes with the predictors. Then we model the conditional variance by a functional regression model applied to the squares of first order model functional errors $\hat{\mathrm{x}}^2 = \left(1 - \hat{m} - \hat{\mathrm{b}}'c\right)^2$, again by means of a GAM/mixed-model representation, obtained by rewriting $o(\cdot)$ and the term $\mathrm{c}(\cdot)'c(\cdot)$, to arrive to an expression similar to (5). This procedure follows the iterative algorithm suggested by Ruppert et al. (2003, p. 264) that we implement as explained in the following subsection.

Model selection can be performed by comparing REML or GCV; indeed since we adopt the mixed model representation, we will use REML by means of the related *AIC* criterion.

3.2 Iterative algorithm for HFRM estimation

To handle heteroskedasticity, we follow the iterative algorithm suggested by Ruppert et al. (2003), which is started by a preliminary standard GAM/mixed-model estimation for the functional mean. Then it is given by

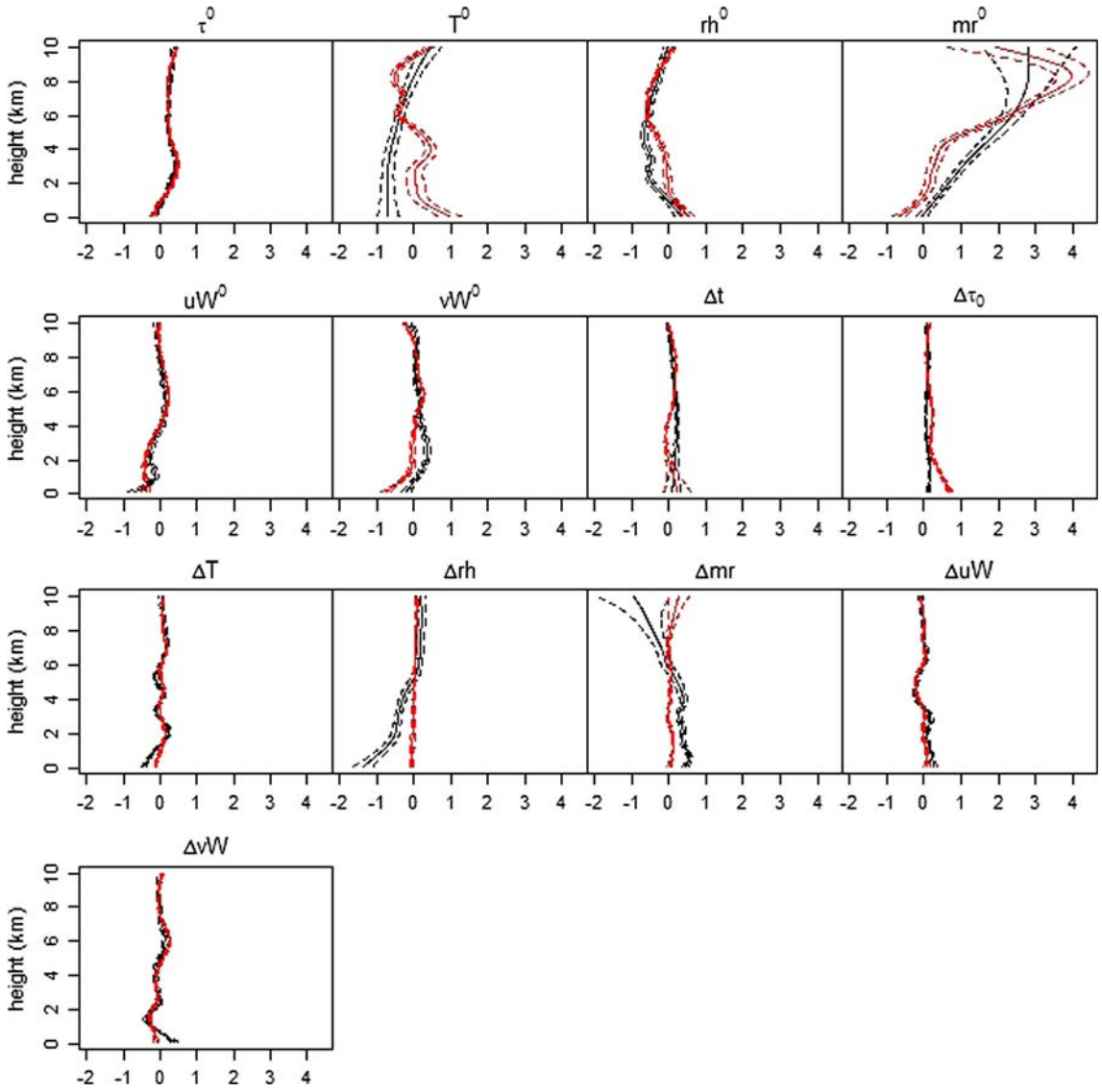

Fig. 8 Initial (*black*) and heteroskedasticity-adjusted (*red*) estimated functional coefficients $\hat{b}_q$ for the functional mean model (3) of pressure collocation mismatch

iterating, up to convergence, the following two steps: a functional regression model estimation step applied to the squared residuals, and an heteroskedastic mixed model estimation step for the functional mean, the latter obtained by pretending that the variance function estimated at the first step is the actual one.

Note that Model (5) can be written in matrix form as $\mathrm{Dl} = \mathrm{An} + \mathrm{x}$, where Dl is the vector of responses and the design matrix A is obtained by stacking all the splines matrices of the individual terms, so that $\mathbf{A} = \underset{\mathrm{n}}{\mathbf{A_{p^0}}}\,\mathbf{A_{Dp}}\,\vec{1}\,\underset{1}{\mathbf{A_1}}\ldots\mathbf{A_Q}$. The model for the variance specified in (4) can be written in a similar matrix form. Following the equivalence between GAM and mixed models, the mean function in (5) can be re-expressed as a mixed model, that is $\mathbf{f} = \mathrm{An} = \mathbf{Xb} + \mathbf{Zu}$ where $\mathbf{X}$ contains the columns of matrix $\mathbf{A}$ corresponding to unpenalized coefficients and $\mathbf{Z}$ contains those columns corresponding to penalized coefficients. Similarly the variance function can be written as $\mathbf{g} = \exp\{\mathrm{Xc} + \mathbf{Zv}\}$. This formulation results in a double-mixed model

$$\mathrm{Dl}\,|\mathbf{u},\mathbf{v} \sim N(\mathrm{Xb} + \mathrm{Zu}, diag\{\exp(\mathrm{Xc} + \mathbf{Zv})\})$$

for which parameter estimation proves to be challenging. Instead, the iterative algorithm provides a way of estimating the parameters (of both the mean and variance functions) that can be easily implemented. The algorithm builds on the fact that if $\mathbf{f}$ is known, then

$$(\mathrm{Dl} - \mathbf{f})^2 \sim Gamma(1/2, 2\exp(\mathbf{Xc} + \mathbf{Zv})).$$

The algorithm steps are defined as follows:

Table 2 Model summary for the functional variance

Parametric coefficients			
	Estimate	Std. error	p-value
c_0	69.967	4.385	\ 2e-16
Smooth terms	edf	F	p-value
$o_0(x^0, y^0, h)$	7.019	35.487	\ 2e-16
$o_D(Dx, Dy, h)$	9.002	37.803	\ 2e-16
s^0	14.804	5.390	4.47e-12
T^0	11.563	31.212	\ 2e-16
rh^0	12.270	9.498	\ 2e-16
mr^0	15.304	8.348	\ 2e-16
uW^0	16.473	12.502	\ 2e-16
vW^0	15.056	9.781	\ 2e-16
Dt	13.095	7.782	\ 2e-16
Ds_0	9.068	67.725	\ 2e-16
DT	9.645	4.048	4.71e-06
Drh	15.824	3.388	1.55e-06
Dmr	4.562	4.985	0.000106
DuW	2.001	5.779	0.003098
DvW	13.701	7.481	\ 2e-16

1. Fit a standard linear mixed model to Dl and get the fitted mean function $\hat{\mathbf{f}}$.
2. Form the squared residuals $\hat{r}^2 = (\text{Dl} - \hat{\mathbf{f}})^2$.
3. Fit to the squared residuals the generalized linear mixed model $Gamma(1/2, 2\exp(\text{Xc} + \mathbf{Zv}))$ and get the fitted function $\exp\hat{\mathbf{g}}$.
4. Fit a heteroskedastic mixed model: pretending that the vector of estimated variance function values, $\hat{\mathbf{g}}$, is the actual variance function, fit the model

$$\text{Dl}\,|\mathbf{u} \sim N(\mathbf{Xb} + \mathbf{Zu}, diag\{\exp(\hat{\mathbf{g}})\}).$$

5. Return to step (2) and iterate.

The algorithm's convergence is determined based on the *AIC*, since the effective degrees of freedom change from iteration to iteration, as well as for the different specification of $m(\cdot)$ and $o(\cdot)$. The algorithm stops when the maximum of the *AIC* rate for $\mathbf{f}$ and $\mathbf{g}$ is smaller than 0.1%, that is

$$max(AICrate_{\mathbf{f}}, AICrate_{\mathbf{g}})\backslash\ \ 0.001$$

where the criterion rate is calculated as

$$AICrate = \frac{AIC^i - AIC^{i-1}}{AIC^{i-1}}$$

and i denotes the ith iteration.

Note that the squared residuals $\hat{\mathbf{r}}^2$ change at each iteration and so the corresponding *AIC* values for $\mathbf{g}$ are not directly comparable. Nevertheless an improvement in modelling $\mathbf{f}$ induces a decay in the *AIC* for $\mathbf{g}$.

4 Collocation uncertainty of atmospheric pressure at Sterling

Data, gathered at discrete and irregularly spaced sampling points, are converted to functional observations through smoothing by using penalized cubic B-splines according to (1) with knots regularly spaced every 50 m and penalty parameter k = 1; these choices allow the observed measurement errors to be small, ensuring that no features of the original data are lost due to oversmoothing and resulting in very small RMSE for all the profiles. Values of smoothed profiles at every 100 m are then considered, so that we recover a common grid with H values for all profiles. In what follows, we consider the difference between coupled profiles at the same height that ranges in 100 – 10000 m. All the analysis is done in R Core Team (2013).

Following the HFRM illustrated in Sect. 3, we model pressure collocation mismatch in Beltsville-Sterling defined as Dpr (in hPa). Figure 4 shows Dpr variability at

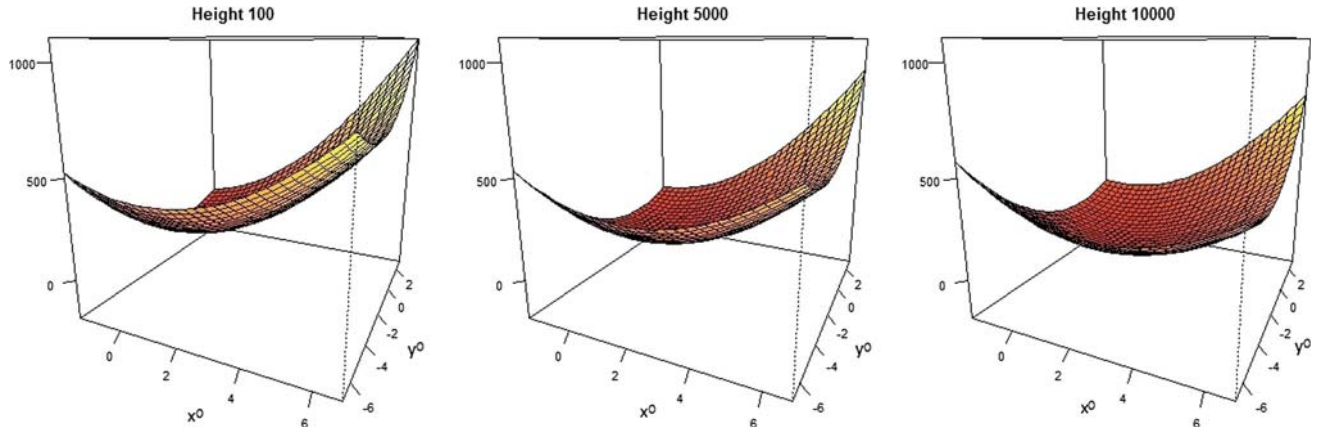

Fig. 9 $\hat{o}_0(x^0, y^0, h)$ for three different values of height

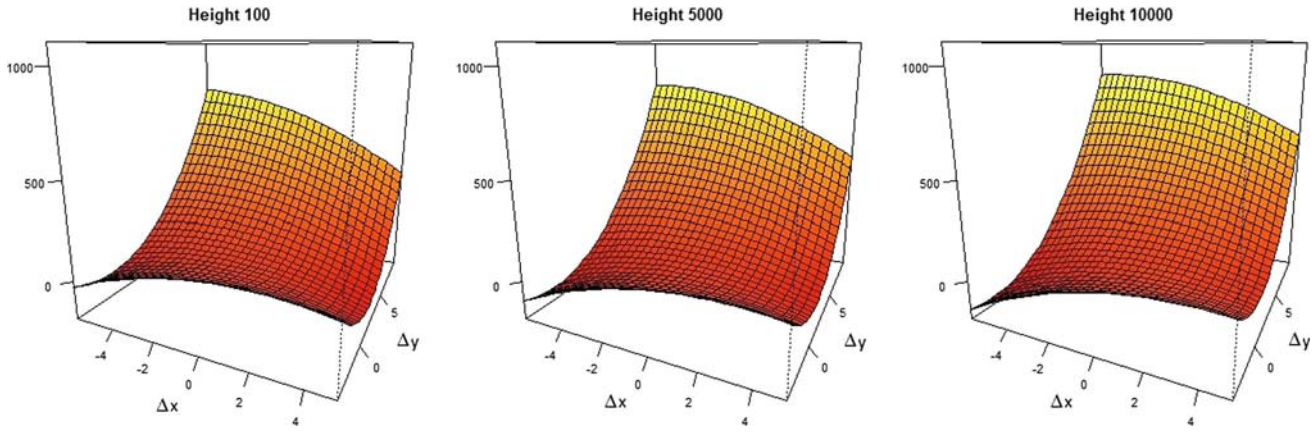

Fig. 10 $\hat{o}_{\mathrm{D}}(\mathrm{D}x, \mathrm{D}y, h)$ for three different values of height

all altitude levels, especially for low value of height. In the HFRM (Eqs. (3) and (4)) we have the following meteorological covariates: temperature (T^0 and DT in K), relative humidity (rh^0 and Drh in %), water vapor mixing ratio (mr^0 and Dmr in g/kg) and orthogonal wind components (uW^0, vW^0, DuW and DvW in m/s) from both collocated radiosondes. To avoid scale effects and facilitate interpretation, the functional covariates $c(\cdot)$ have been standardized so that the total profile average is zero and the total profile variance is unity. Note that differences are taken at the same height value but with a mismatch in time (less than 3 h). Hence, we need to include further covariates in the model to adjust for the change in local meteorological conditions within the time period in between matched launches. A further source of variability comes from the fact that the 32 pairs of launches are distributed in a three year period and hence they may have been launched in different seasons and time of the day. Thus we consider measurement calendar time (s^0 and Ds_0) and flight duration difference (Dt in seconds). In particular, Ds_0 represents the difference in measurement calendar time at launch, and thus it is smaller than 3 h. We consider this (scalar) covariate instead of Ds because of the relationship $\mathrm{Ds} = \mathrm{Ds}_0 + \mathrm{D}t$.

All three alternatives for $m(p^0, \mathrm{D}p)$ in (3) and $o(p^0, \mathrm{D}p)$ in (4), namely (A), (B) and (C), are considered. Convergence results from the iterative algorithm detailed in Sect. 3.2 are summarized in Fig. 5. According to the *AIC* criterion for **f**, Model (A) is considered to be the best since $AIC = 770.62, 921.01$ and 852.39 for (A), (B) and (C) respectively. Indeed, AIC values for Model (A) are systematically smaller than those for (B) and (C) with increasing number of iterations. As already explained in Sect. 3.2, *AIC* values for **g** are not directly comparable, but Fig. 5 shows the expected decay. Moreover, we can see that *AIC* decays faster for Model (A) than for (B) and (C) when heteroskedasticity is taken into account.

The computational time is also smaller for Model (A), 5.85 hours, than for the other two models; although there is not much difference with respect to Model (B), converging in 6.03 hours, running time is considerably smaller than for Model (C) that takes 13.3 h.

From now onwards, reported results are based on Model (A) where the coordinates do not act independently, but their interaction is allowed by including two smooth functions of p^0 and Dp.

The final fitted model for the functional mean is summarized in Table 1. The intercept b_0 is assumed to be independent of height so that $\hat{\mathrm{b}_0} = 0.89486$ can be seen as the overall collocation bias between Beltsville and Sterling radiosoundings. All smooth functions of the covariates were found to be significant according to the zero-effect Wald-type test (Wood 2013) of smooth components in GAMs as implemented in mgcv (Wood 2012), as shown by the p-values in Table 1. Moreover, the effective degrees of freedom (edf) indicate nonlinear effects of all the considered covariates, with the exception of Drh whose $\mathrm{b}(h)$ appears almost linear (edf = 2).

Figures 6 and 7 show the estimated $\hat{m}_0$ and $\hat{m}_{\mathrm{D}}$ for three different values of h. Since they are 3D smooth components, we can only visualize them by fixing one of the three variables (height in this case). From Fig. 6, it seems clear that a model where the effect of longitude, latitude and height is merely additive, as Model (C) states, may not be very appropriate, given that the shape of the surface changes considerably depending on the value of height. The same is true for $\hat{m}_{\mathrm{D}}$ (Fig. 7), although in this case the difference is less evident than for $\hat{m}_0$.

The estimated functional coefficients are plotted in Fig. 8 along with 95 % confidence bands and show the influence of each of the covariates on the collocation drift. Estimated coefficients adjusted for heteroskedasticity (i.e. at the end of the iteration process) are shown in red, while initial estimates are shown in black. By incorporating the

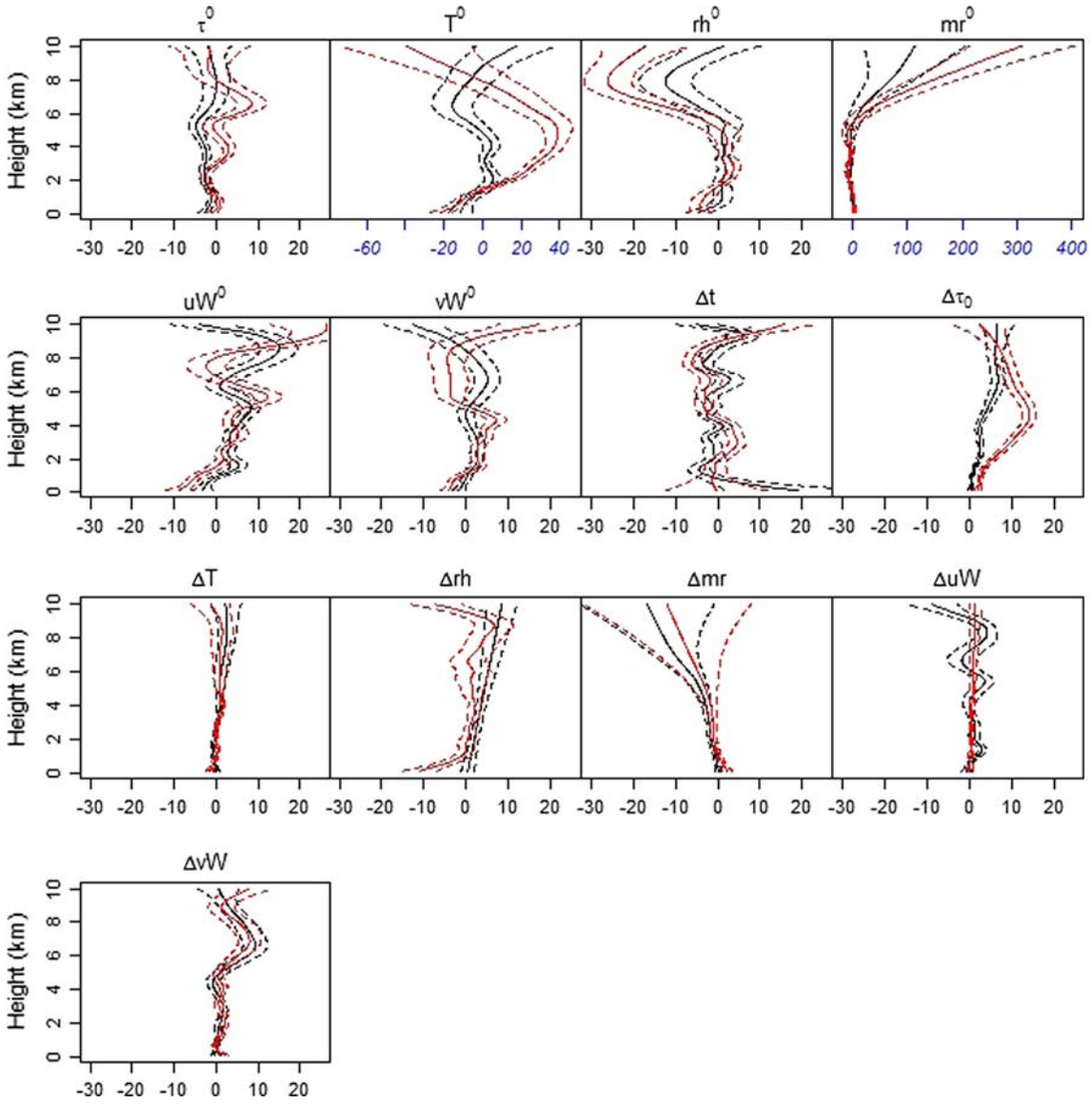

Fig. 11 Initial (*black*) and heteroskedasticity-adjusted (*red*) estimated functional coefficients $\hat{c}_q$ for the functional variance model (4) of pressure collocation mismatch. Note that the axis in *blue* (T^0 and mr^0) are on a different scale and that the *black* and *red* estimates are not directly comparable

heteroskedasticity, the 95 % confidence bands become generally narrower and the functional coefficients associated to meteorogical covariates change in shape and magnitude, especially for T^0 and mr^0, while those of time related covariates appear to change less than the former.

With an adjusted determination coefficient $R^2 = 0.952$, the model summarized in Table 1 misses only 4.8 % of the collocation uncertainty which is covered by $\mathrm{r}^2_{\mathrm{x}}(\cdot)$. The latter is estimated by the functional log-linear model applied to the squares of first order model functional errors $\hat{\mathrm{x}}^2 = \left[1 - \hat{\mathrm{b}}' c\right]^2$, according to (4). The corresponding fitted model is summarized in Table 2, where *p*-values indicate that the collocation 2nd order uncertainty of pressure depends on the same covariates as the collocation drift. In addition, the effective degrees of freedom (edf) support the nonlinearity of the functional coefficients $\mathrm{c}(h)$; only D*uW* has an almost linear effect with edf = 2.001.

Figures 9 and 10 show the estimated $\hat{\vartheta}_0$ and $\hat{\vartheta}_{\mathrm{D}}$ for three different values of *h*. The term $\hat{\vartheta}_0$ becomes flatter as height increases, while the term $\hat{\vartheta}_{\mathrm{D}}$ remains very similar for different values of height. The estimated functional coefficients are plotted in Fig. 11 along with 95% confidence bands and show the influence of each of the covariates on the skedastic term. Both initial and final estimates of the iterative algorithm are shown in black and red,

respectively. However, as already said, the squared residuals $\mathbf{f}^2$ change at each iteration and so the corresponding estimated functional coefficients are not directly comparable. The effects' magnitude is much larger for mr^0 whose functional coefficient increases abruptly after $h = 6$ km. Also T^0's coefficient has a larger range than the remaining covariates (see Fig. 11).

5 Conclusions

In this paper we presented a new model for 3D functional data based on the extension of a previously introduced unidimensional heteroskedastic regression model. Fitting follows from a GAM/mixed-model representation. In fact, this reformulation as a double mixed model, together with the implementation of an iterative algorithm optimizing an AIC criterion, allows us to handle the impact of covariates on conditional uncertainty by means of functional heteroskedasticity.

The model describes both conditional mean and variance as a sum of a 3D functional term and some unidimensional functional regression components. This results in great flexibility as shown by the application to collocation uncertainty of atmospheric thermodynamic profiles.

In particular, considering collocation uncertainty of atmospheric pressure, the new 3D component is shown to improve model fitting with respect to the purely undimensional model which was introduced by Fassò et al. (2013) in the frame of collocation uncertainty of relative humidity. Moreover, the iterative algorithm allows to adjust model estimates of collocation drift in the presence of heteroskedasticity.

The difference in pressure actually measured by couples of weather balloons is known not to match exactly the barometric formula but to require corrections for variations in density and meteorological variables such as temperature, humidity and wind conditions, see e.g. Berberan-Santos et al. (2010). It is interesting to note that the model obtained in this paper, with a fitting of $R^2 = 0.95$ and satisfying AIC parsimony criterion, also includes a number of terms that take into account time and space for the two collocated measurements. Moreover, it shows that these effects are not linear since they are smoothly changing in shape along vertical direction and horizontal distance. In addition, the small unexplained collocation uncertainty changes in magnitude as explained by the heteroskedastic 3D component.

Acknowledgments The authors would like to thank Belay Demoz and Fabio Madonna for useful discussions and comments and two anonymous referees whose comments and suggestions improved the reading and quality of the manuscript.

References

Berberan-Santos MN, Bodunov EN, Pogliani L (2010) On the barometric formula inside the earth. J Math Chem 47(3):990–1004

Caballero W, Giraldo R, Mateu J (2013) A universal kriging approach for spatial functional data. Stoch Environ Res Risk Assess 27(7):1553–1563

Cuevas A (2014) A partial overview of the theory of statistics with functional data. J Stat Plan Inference 147:1–23

Eilers PHC, Marx BD (2003) Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. Chemom Intell Lab Syst 66:159–174

Fassò A, Ignaccolo R, Madonna F, Demoz B (2013) Statistical modelling of atmospheric vertical profiles and the collocation problem. Atmos Meas Tech Discuss 6:7505–7533. doi:10.5194/amtd-6-7505-2013

Escabias M, Valderrama J, Aguilera AM, Santofimia ME, Aguilera-Morillo MC (2013) Stepwise selection of functional covariates in forecasting peak levels of olive pollen. Stoch Environ Res Risk Assess 27(2):367–376

Ferraty F, Vieu P (2006) Nonparametric functional data analysis: theory and practice. Springer, New York

Gijbels I, Prosdocimi I, Claeskens G (2010) Nonparametric estimation of mean and dispersion functions in extended generalized linear models. Test 19:580–608

Guo W (2004) Functional data analysis in longitudinal settings using smoothing splines. Stat Methods Med Res 13:49–62

Harezlak J, Coull BA, Laird NM, Magari SR, Christiani DC (2007) Penalized solutions to functional regression problems. Comput Stat Data Anal 99:4911–4925

Hastie T, Tibshirani R (1993) Varying-coefficient models. J R Stat Soc B 55:757–796

Horvàth L, Kokoszka P (2012) Inference for functional data with applications. Springer, New York

Ignaccolo R, Mateu J, Giraldo R (2013) Kriging with external drift for functional data for air quality monitoring. Stoch Environ Res Risk Assess. doi:10.1007/s00477-013-0806-y

Immler FJ, Dykema J, Gardiner T, Whiteman DN, Thorne PW, Vomel H (2010) Reference quality upper-air measurements: guidance for developing GRUAN data products. Atmos Meas Tech 3:1217–1231

Ivanescu AE, Staicu AM, Greven S, Scheipl F, Crainiceanu CM (2012) Penalized function-on-function regression (April 2012). Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 240

Karlis D, Vasdekis VGS, Banti M (2009) Heteroscedastic semiparametric models for domestic water consumption aggregated data. Environ Ecol Stat 16:355–367

Nash J, Oakley T, Vömel H, LI Wei (2010) WMO Intercomparison of high Quality Radiosonde Systems Yangjiang, China, 12 July–3 August 2010; WMO report reference number IOM 107 (TD 1580). available at: http://www.wmo.int/pages/prog/www/IMOP/publications-IOM-series.html

Ngo L, Wand MP (2004) Smoothing with mixed model software. J Stat Softw 71(9):1–54

Nott DJ (2006) Semiparametric estimation of mean and variance functions for non-Gaussian data. Comput Stat 21:603–620

Ramsay JO, Silverman BW (2005) Functional data analysis. Springer, New York

R Core Team (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Reiss PT, Ogden RT (2009) Smoothing parameter selection for a class of semiparametric linear models. J R Stati Soc B 71:50517523

Robinson GK (1991) That BLUP is a good thing: the estimation of random effects. Stati Sci 6:15–32

Ruiz-Medina MD, Espejo RM (2012) Spatial autoregressive functional plug-in prediction of ocean surface temperature. Stoch Environ Res Risk Assess 26:335–344

Ruppert D, Wand MP, Carroll RJ (2003) Semiparametric regression. Cambridge University Press, New York

Seidel DJ, Sun B, Pettey M, Reale A (2011), Global radiosonde balloon drift statistics. J Geophys Res 116:D7

Speed T (1991) Comment on paper by Robinson. Stati Sci 6:421744

Thorne PW, Vömel H, Bodeker G et al (2013) GCOS reference upper air network (GRUAN): Steps towards assuring future climate records. AIP Conference Proceedings 1552:1042–1047. doi:http://dx.doi.org/10.1063/1.4821421

Wahba G (1990) Spline models for observational data. SIAM, Philadelphia

Wand MP (2003) Smoothing and mixed models. Comput Stat 18:223–249

Wang H, Akritas MG (2010) Inference from heteroscedastic functional data. J Nonparametric Stat 22(2):149–168

Wood SN (2004) Stable and efficient multiple smoothing parameter estimation for generalized additive models. J Am Stat Assoc 99:673–686

Wood AN (2006) Generalized additive models: an introduction with R. Chapman & Hall/CRC, Boca Raton

Wood SN (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J R Stat Soc B 73(1):3–36

Wood SN (2012) mgcv: Mixed GAM computation vehicle with GCV/AIC/REML smoothness estimation, R package version 1.7–22

Wood SN (2013) On *p* values for smooth components of an extended generalized additive model. Biometrika 100(1):221–228

Zhang JT (2013) Analysis of variance for functional data. Chapman & Hall/CRC, Boca Raton