ELSEVIER

Physica D 151 (2001) 142–174

PHYSICA D

www.elsevier.com/locate/physd

The Kalman–Lévy filter

Didier Sornette*, Kayo Ide [c]

[a] *Department of Earth and Space Sciences, Institute of Geophysics and Planetary Physics, University of California at Los Angeles, Los Angeles, CA 90095-1567, USA*

[b] *Laboratoire de Physique de la Matière Condensée, CNRS UMR 6622 and Université de Nice-Sophia Antipolis, 06108 Nice Cedex 2, France*

[c] *Department of Atmospheric Sciences, Institute of Geophysics and Planetary Physics, University of California at Los Angeles, Los Angeles, CA 90095-1565, USA*

Received 28 April 2000; received in revised form 8 December 2000; accepted 16 January 2001
Communicated by U. Frisch

Abstract

The Kalman filter combines forecasts and new observations to obtain an estimation which is optimal in the sense of a minimum average quadratic error. The Kalman filter has two main restrictions: (i) the dynamical system is assumed linear and (ii) forecasting errors and observational noises are projected onto Gaussian distributions. Here, we offer an important generalization to the case where errors and noises have heavy tail distributions such as power laws and Lévy laws. The main tool needed to solve this "Kalman–Lévy" filter is the "tail-covariance" matrix which generalizes the covariance matrix in the case where it is mathematically ill-defined (i.e. for power law tail exponents $\mu \leq 2$). We present the general solution and discuss its properties on pedagogical examples. The standard Kalman–Gaussian filter is recovered for the case $\mu = 2$. The optimal Kalman–Lévy filter is found to deviate substantially from the standard Kalman–Gaussian filter as μ deviates from 2. As μ decreases, the filter favors more strongly the better one of the forecast and the observation, based on the tail-covariance matrix because a small exponent μ implies large errors with significant probabilities. In terms of implementation, the price-to-pay associated with the presence of heavy tail noise distributions is that the standard linear formalism valid for the Gaussian case is transformed into a nonlinear matrix equation for the Kalman–Lévy filter. Direct numerical experiments in the univariate case confirms our theoretical predictions. © 2001 Published by Elsevier Science B.V.

Keywords: Kalman–Lévy filter; Non-normal noise distribution; Tail-covariance matrix; Data assimilation; Optimal estimation

1. Introduction and motivation

The Kalman filter provides the optimal resolution of the problem of data assimilation under the hypothesis that the system observables evolve according to linear maps, are linearly related to the true variables and that the noise acting on the true dynamics and the measurement errors are mutually uncorrelated and Gaussian.

Two main limitations restrict the performance of the Kalman filter:

- the nonlinearity of the real system dynamics;
- the non-normality of the noises.

* Corresponding author.
E-mail address: didier.sornette@unice.fr (D. Sornette).

0167-2789/01/$ – see front matter © 2001 Published by Elsevier Science B.V.
PII: S0167-2789(01)00228-7

The first item has been addressed partly by using so-called "extended" Kalman filters that amount essentially to perform local linearizations [18].

With respect to the non-normality of the noises, the general condition for using the Kalman filter is that their covariance functions exist, which is satisfied for noise density distributions decaying faster than $1/\omega^3$. When the noise distributions are not Gaussian, the validity of the Kalman filter relies on the existence of a central limit theorem for state estimators, which exists when the random terms in the model have arbitrary distribution with tail decaying faster than $1/\omega^3$ [25]. For practical applications, the existence of the central limit theorem does not suffice as a finite observation time may lead to large deviations from the asymptotic results [24]. The knowledge of convergence rates in the central limit theorem are then necessary for the development of tests of the validity of the model [2].

Our purpose here is to extend these results to the regime where the existence of the covariance function is not warranted as occurs for Lévy distributions of noises, or when the covariance functions do exist but the convergence to the asymptotic result given by the central limit theorem is extremely slow making the asymptotic result useless in practice, as occurs for power law distributions [24]. In order to illustrate this idea, consider the sum of N identically independent distributed random variables, with a power law probability density function $1/\omega^{1+\mu}$ with exponent $\mu > 2$. Since the variance σ^2 is finite, the central limit theorem applies and the distribution of the sum converges to the Gaussian law with standard deviation $\approx \sigma\sqrt{N}$. For N finite, the Gaussian law only describes the central part of the distribution of the sum, up to a cross-over $S_0 \approx \sigma\sqrt{N \ln N}$ [24]. Beyond S_0, the distribution of the sum is a power law with the same exponent μ and the weight in probability of this power law tail decays slowly. For instance, for $\mu = 3$, the weight in probability of the power law tail decays as $\propto 1/(\sqrt{N}\ln^{3/2} N)$ as N increases. In practice, consider the three sample sizes $N = 10^2, 10^4$ and 10^6. The corresponding cross-over values are, respectively, $S_0 \approx 2.1$, 3 and 3.7 times the standard deviation of the central Gaussian part of the distribution of the sum. These estimates suggest that even if the general condition for using the Kalman filter is satisfied for noise density distributions decaying faster than $1/\omega^3$, the Gaussian (or covariance) approach, which is optimal in the linear least-variance estimation, is not necessarily optimal when relatively large fluctuations (of size equal to two standard deviations or larger) occur.

We thus propose to explore how the optimal Kalman filter is modified when the objective is to minimize, not the variance of the error estimation but a natural measure of the large errors, namely the tail of the distribution of the Euclidean norm of the errors. Our approach thus develops an alternative class of linear unbiased estimators different from the standard linear least-variance estimation. Our emphasis is in trying to control and minimize the large (and rare) errors, with the penalty that the variance, if it exists, will be sub-optimal compared to the standard linear least-variance estimation. In this goal, we focus our attention on noises in the observable and the unobservable variables both distributed with a power law tail of fixed exponent μ and known amplitudes (scale factors). This assumption offers a well-defined theoretical limit of "large fluctuations", putting the emphasis on the complement to the "small noise" limit captured by the standard linear least-variance approach.

In addition to its pure theoretical interest, we observe that many systems in Nature are claimed to exhibit power law distributions [24] and the present results may thus have direct application. Power law distributions have been found to quantify the size–frequency Gutenberg–Richter distribution of earthquakes, of hurricanes [20], of volcanic eruptions, of floods, of meteorite sizes and so on. The distribution of seismic fault lengths is also documented to be a power law with exponent $\mu \simeq 1$. In the insurance business, recent studies has shown that the distribution of losses due to business interruption resulting from accidents [26,27] is also a power law with $\mu \simeq 1$.

Several previous works have addressed related issues. Le Breton and Musiela's work [16] is closest to ours but has limitations: (i) it is based on a continuous time description of the dynamical and observation processes and is thus more difficult to apply to concrete situations; (ii) it minimizes the difference between the true dynamics and a filtered observation in the L^μ-norm sense where μ is the index exponent of the Lévy distributions; in other words it relies on an explicit solution of the dynamics; in this way, Le Breton and Musiela circumvent the two delicate

questions of generalizing the covariance matrix and of choosing the objective function to optimize; (iii) it does not really address the genuine Kalman problem which consists in *mixing* forecasts and observations. Recently, Ahn and Feldman [1] have proposed a filter for the case where the signal is Gaussian while the observation noise is a Lévy process. Their filter is optimal in the sense of minimizing the L^2 error, i.e. the distance between the true dynamics and the filter output. This choice of the L^2 is realizable in their case because the signal is assumed Gaussian and the integrability assumption is thus satisfied. The problem, however, is that the optimal filter will depend in general upon the choice of the norm that measures the prediction error. In addition, due to the rather intractable nonlinear recursive solution they obtain, they propose a sub-optimal filter for numerical purpose.

Our approach circumvents these difficulties by focusing on large errors. Our goal is thus to minimize the large errors between the analysis (obtained by assimilation of observation with prediction) and the true trajectory. With this choice, the technical problem we have to solve is to characterize the tails of error distributions, which are a function of the tail of the distributions of individual dynamical variables and of noises and of their mutual dependence. In order to determine these dependencies, we propose to use the concept of a "tail-covariance" matrix that generalizes the standard covariance matrix of the analyzed signal to the case of power laws and Lévy distributions. Tail-covariance matrices are constructed as the matrices of *signed* scale factors (the scale factors being defined as the global amplitudes of the power law tails) of the distribution of all the products of the system and forecasted variables. The main idea is thus to replace the characterization of errors and of correlations by the ensemble of distributions of the products of all possible pairs of variables. Our approach thus replaces a reasoning based on the second centered moments into one based on the tails of the distribution of deviations from true values. It is shown that the natural error amplitude to be minimized is the trace of the tail-covariance, which is a straightforward extension of the standard approach which determines the Kalman filter by minimizing the average errors quantified by the trace of the covariance matrix of the analyzed signal.

This paper is constructed in a pedagogical manner, from the simple univariate filter problem to the full multivariate Kalman filter. In Section 2, we formulate the problem for the univariate filter with power law noises, corresponding to the situation where only novel observations are assimilated without mixing with a dynamical forecast. A detailed discussion is offered to compare the power law case with the standard Gaussian case. In Section 3, we generalize this problem to the case of multivariate estimators. In Section 4, we address the Kalman problem of mixing both forecasts and observations with power law and Lévy law errors and develop our general solution. The special univariate case is studied in great detail by numerical experiments to contrast the performances of the Kalman–Lévy (KL) with those of the standard Kalman–Gaussian (KG) filter.

2. Univariate estimator: principle of estimation with power laws

2.1. Formulation of the problem

Take two samples x^{f} and x^{o} of an observable state variable x; the superscripts $\{\cdot\}^{\mathrm{f,o}}$ correspond to forecast and observation of the data assimilation system, respectively. The two samples are contaminated by independent noise ω^{f} and ω^{o}. The estimate $\hat{x}$ of x is sought as a linear combination of x^{o} and x^{f} with the corresponding *positive* weights K^{f} and K^{o}

$$\hat{x} = K^{\mathrm{f}} x^{\mathrm{f}} + K^{\mathrm{o}} x^{\mathrm{o}}. \tag{1}$$

One of our main goals is to determine the optimal weights so as to minimize the resulting uncertainty of $\hat{x}$. We make two assumptions concerning samples in this study.

The first assumption that the two samples are unbiased

$$\langle x^{\mathrm{f}} \rangle = \langle x^{\mathrm{o}} \rangle = \langle x \rangle, \tag{2}$$

where $\langle z\rangle$ is the expectation of z. Expectation of the state variable $\langle x\rangle$ is not known a priori. By requiring further that the estimate should be unbiased $\langle\hat{x}\rangle = \langle x\rangle$, we obtain a relation between the two weights $K^{\mathrm{f}} + K^{\mathrm{o}} = 1$ which allows us to rewrite (1) as

$$\hat{x} = x^{\mathrm{f}} + K^{\mathrm{o}}(x^{\mathrm{o}} - x^{\mathrm{f}}), \tag{3}$$

and therefore reduces the problem to the determination of a single unknown $0 \leq K^{\mathrm{o}} \leq 1$. This expression (3) can be interpreted as a filtering of the observed data x^{o} into the dynamical forecast x^{f} by a weighted increment $K^{\mathrm{o}}(x^{\mathrm{o}} - x^{\mathrm{f}})$.

The second assumption is that both sample errors

$$\omega^{\mathrm{f,o}} = x^{\mathrm{f,o}} - x, \tag{4}$$

are distributed according to a power law as defined in (5), or according to a Lévy law as defined in Appendix A, with the same exponent $\mu^{\mathrm{f,o}} \equiv \mu$. The important property for our purpose is that the tails of the probability density functions of the two sample errors are independent and given by

$$P(\omega^{\mathrm{f,o}}) \simeq \frac{C_{\pm}^{\mathrm{f,o}}}{|\omega^{\mathrm{f,o}}|^{1+\mu}}, \quad |\omega^{\mathrm{f,o}}| \to \pm\infty, \tag{5}$$

where the subscript $\{\cdot\}_{\pm}$ reflects that the tail distribution can be asymmetric depending on the sign of $\omega^{\mathrm{f,o}}$. In this study, we focus on the symmetric case, i.e., $C_{\pm}^{\mathrm{f,o}} \equiv C^{\mathrm{f,o}}$.

The family of power laws is characterized by two parameters, the exponent μ and the 'scale factor' C. The exponent μ, on one hand, controls the decay rate of the probability as well as its scaling (or self-similar) properties. The scale factor C, on the other hand, controls the overall amplitude of the power law tail, i.e., the larger it is, the more important is the power law tail. More precisely, if the power law tail (5) holds for ω larger than some minimum value $\omega_{\mathrm{min}}^{\mathrm{f,o}}$, the weight in probability of the power law, i.e. the probability that ω is larger than $\omega_{\mathrm{min}}^{\mathrm{f,o}}$ is $(C^{\mathrm{f,o}}/\mu)(\omega_{\mathrm{min}}^{\mathrm{f,o}})^{-\mu}$. As shown in Appendix A in the case of Lévy laws, the scale parameter fully characterizes the distribution for all variations (and not only in the tail).

Notice that (5) can be rewritten in terms of the dimensionless variable $\omega/C^{1/\mu}$ with the superscripts dropped for simplicity

$$P(\omega)\,\mathrm{d}\omega \simeq \frac{1}{|\omega/C^{1/\mu}|^{1+\mu}}\,\mathrm{d}\left(\frac{\omega}{C^{1/\mu}}\right), \quad |\omega| \to \pm\infty, \tag{6}$$

showing that $C^{1/\mu}$ is the characteristic scale of the self-similar fluctuations of ω. For $\mu \leq 2$ (resp. $\mu \leq 1$), the variance (resp. mean) is not defined mathematically. We shall see the effect of these μ-dependent properties throughout this study.

Using the continuation property given in [7] (Problem 15 of Section 12) on the characteristic function of distributions with power tails, we obtain the following results. Let us call a μ-variable a variable with a distribution function with a power law tail. Then, we have the following:

1. If ω_i and ω_j are two independent μ-variables characterized by the scale factors $C_i^{\pm}$ and $C_j^{\pm}$, then $\omega_i + \omega_j$ is also a μ-variable with $C^{\pm}$ given by $C_i^{\pm} + C_j^{\pm}$.
2. If ω is a μ-variable with scale factor C, then $p \times \omega$ (where p is a real number) is a μ-variable with scale factor $p^{\mu}C$. If $p < 0$ and the distribution of ω is symmetric, then $p \times \omega$ is a μ-variable with scale factor $|p|^{\mu}C$.
3. If ω is a μ-variable, then $\mathrm{sign}(\omega)|\omega|^q$, with $q > 0$, is a μ/q-variable.
4. If ω_i and ω_j are two independent μ-variables, then the product $x = \omega_i\omega_j$ is also a μ-variable up to logarithmic corrections.

Using the rules (1) and (2), we find that the distribution of $\hat{x}$ is also a power law (5) with the same exponent μ, but with an adjusted scale factor $\hat{C}$ given by

$$\hat{C} = (1 - K^{\mathrm{o}})^{\mu} C^{\mathrm{f}} + (K^{\mathrm{o}})^{\mu} C^{\mathrm{o}}. \tag{7}$$

The expression (7), which is valid for $0 \leq K^{\mathrm{o}} \leq 1$ such that the scale factors remain positive, can be reduced to the usual result for the Gaussian distributions [8]

$$\hat{\sigma}^2 = (1 - K^{\mathrm{o}})^2 (\sigma^{\mathrm{f}})^2 + (K^{\mathrm{o}})^2 (\sigma^{\mathrm{o}})^2, \tag{8}$$

by setting the exponent $\mu = 2$ and replacing the scale factors $C^{\mathrm{f,o}}$ with $(\sigma^{\mathrm{f,o}})^2$. We thus see that the scale parameter C is the generalization of the variance σ^2. The technical reason for this comes from the form of the characteristic function of a distribution with a power law tail as given in [7] (Problem 15 of Section 12). The situation is even simpler to discuss when considering symmetric Lévy laws whose characteristic functions read

$$\hat{L}_{\mu}(k) = \exp(-a_{\mu}|k|^{\mu}) \quad \text{for } 0 < \mu \leq 2, \tag{9}$$

where a_{μ} is a constant proportional to the scale parameter C [10]. The important point is that for μ strictly less than 2, the inverse Fourier transform of $\hat{L}_{\mu}(k)$ gives a power law tail, while for $\mu = 2$, it gives a Gaussian law. The continuity between the expressions (7) and (8) can thus be traced back to that of $\hat{L}_{\mu}(k)$ as a function of μ at $\mu = 2$ (see Appendix A for a more formal derivation of this fact).

2.2. Solution

The standard optimal estimation methodology consists in minimizing the variance $\hat{\sigma}^2$ with respect to the weight factor K^{o}. The solution for K^{o} then gives the best weighting in the sense that we remain with the smallest uncertainty from the estimation with the novel data, by minimizing the expectation distance between $\hat{x}$ and x in the mean-square sense. In order to generalize this methodology to the situation where the errors are distributed according to power law distributions, we propose the following central idea, i.e., to minimize the scale factor $\hat{C}$ with respect to K^{o}

$$\frac{\partial}{\partial K^{\mathrm{o}}} \hat{C} = \mu[-(1 - K^{\mathrm{o}})^{\mu-1} C^{\mathrm{f}} + (K^{\mathrm{o}})^{\mu-1} C^{\mathrm{o}}] = 0, \tag{10}$$

$$\frac{\partial^2}{\partial (K^{\mathrm{o}})^2} \hat{C} = \mu(\mu - 1)[(1 - K^{\mathrm{o}})^{\mu-2} C^{\mathrm{f}} + (K^{\mathrm{o}})^{\mu-2} C^{\mathrm{o}}] \geq 0. \tag{11}$$

The justification for this procedure is that the uncertainty in the estimation $\hat{x}$ of x is inescapably distributed according to a power law distribution with the same exponent μ as a result of the rules (1) and (2) given above. Consequently, the optimization using the weight K^{o} can be performed only in one purpose, namely to decrease the global amplitude of the power law controlled by $\hat{C}$, but without be able to distort its shape defined by μ. The optimal weight K^{o} as the solution to (10) depends on the value of the exponent μ as follows:

1. For $\mu > 1$, the minimization of $\hat{C}$ given by (7) with respect to K^{o} gives

$$K^{\mathrm{o}} = \frac{1}{1 + \lambda^{\mu/(\mu-1)}}, \tag{12}$$

where

$$\lambda \equiv \frac{(C^{\mathrm{o}})^{1/\mu}}{(C^{\mathrm{f}})^{1/\mu}} \tag{13}$$

is the ratio of the characteristic error size of the two samples as defined in the distribution (6). The resulting optimal scale factor is

$$C^{\mathrm{a}} = [\lambda^{\mu}/(1+\lambda^{\mu/(\mu-1)})^{\mu-1}]C^{\mathrm{f}}, \tag{14}$$

where the superscript $\{\cdot\}^{\mathrm{a}}$ stands for analysis according to the data assimilation convention.

2. For $\mu < 1$, there is no optimal solution for K^{o} which is not on the boundaries of the search interval and that minimizes $\hat{C}$, because $\partial^2\hat{C}/\partial(K^{\mathrm{o}})^2 < 0$ violates the second condition in (11). Physically this implies that the fluctuations are so wild that the estimation by the weighted average is not a good strategy, and that only the measurement with the smallest scale factor should be kept for the estimation of $\hat{x}$

$$K^{\mathrm{o}} = \begin{cases} 0 & \text{for } C^{\mathrm{o}} > C^{\mathrm{f}}, \\ 1 & \text{for } C^{\mathrm{o}} < C^{\mathrm{f}}, \end{cases} \tag{15}$$

and therefore

$$C^{\mathrm{a}} = \begin{cases} C^{\mathrm{f}} & \text{for } C^{\mathrm{o}} > C^{\mathrm{f}}, \\ C^{\mathrm{o}} & \text{for } C^{\mathrm{o}} < C^{\mathrm{f}}. \end{cases} \tag{16}$$

This second case $K^{\mathrm{o}} = 1$, consisting in trusting observations over the forecast, is known as "direct substitution" [6]. Here, we have shown that it constitutes indeed the best strategy in the specific case $\mu \leq 1$ and $C^{\mathrm{o}} < C^{\mathrm{f}}$. Note that a similar solution applies when the two observations have different exponent μ: full weight $K^{\mathrm{o}} = 0$ or 1 should be put on the observation with the largest exponent, as it has the smallest fluctuations. This solves the general case as well.

To make the problem interesting, in this paper, we consider all noise sources to have the same exponent μ, so that the problem is a "fight between scale-factors". This case is not as restricted as it would appear at first site: if the mechanisms leading to the power law tail are intertwined, such as for instance with a common source of underlying multiplicative noise, it can be shown [11,15] that the power law exponent of the different variables will be the same as soon as there is non-vanishing coupling between the variables. This case will be investigated in a forthcoming work.

The following argument retrieves (15). When neither the variance nor the mean exist, and when the minimization (10) of the scale factor becomes meaningless, the last natural quantity to estimate is the probability that the error remaining after assimilation is smaller than the error on the two measurements. Suppose that we have the knowledge that the errors in the second measurement are larger in probability than that of the first measurement, i.e. $C^{\mathrm{o}} > C^{\mathrm{f}}$. We then require the maximization of the probability $P_{\mathrm{improvement}}$ that

$$|(1-K^{\mathrm{o}})\omega^{\mathrm{f}} + K^{\mathrm{o}}\omega^{\mathrm{o}}| \leq |\omega^{\mathrm{f}}|. \tag{17}$$

Let us assume that ω^{f} is found positive. Then, this probability is the same as the probability that

$$\left(-\frac{2}{K^{\mathrm{o}}}+1\right)\omega^{\mathrm{f}} \leq \omega^{\mathrm{o}} \leq \omega^{\mathrm{f}}. \tag{18}$$

The probability for (17) to be verified is thus

$$P_{\mathrm{improvement}} = 2\int_0^{\infty} \mathrm{d}\omega^{\mathrm{f}}\, P^{\mathrm{f}}(\omega^{\mathrm{f}}) \int_{((-2/K^{\mathrm{o}})+1)\omega^{\mathrm{f}}}^{\omega^{\mathrm{f}}} P^{\mathrm{o}}(\omega^{\mathrm{o}})\,\mathrm{d}\omega^{\mathrm{o}}, \tag{19}$$

where the factor 2 comes from the counting of the cases where ω^{f} can be found negative. By taking the derivative of (19) with respect to K^{o}, we obtain

$$\frac{\mathrm{d}P_{\text{improvement}}}{\mathrm{d}K^{\mathrm{o}}} = -\frac{4}{(K^{\mathrm{o}})^2}\int_0^{\infty} \mathrm{d}\omega^{\mathrm{f}}\, \omega^{\mathrm{f}} P^{\mathrm{f}}(\omega^{\mathrm{f}}) P^{\mathrm{o}}\left(\left(-\frac{2}{K^{\mathrm{o}}}+1\right)\omega^{\mathrm{f}}\right), \tag{20}$$

which is always negative. Thus, the probability that the error is reduced is maximum for $K^{\mathrm{o}} = 0$, i.e. without assimilating the new observation. Intuitively, the power law tails with exponent $\mu < 1$ are so "wild" that it is preferable to keep only the observation with the smallest scale factor. A similar derivation holds in the case where the errors in the second measurement are smaller in probability than that of the first measurement, i.e. $C^{\mathrm{o}} < C^{\mathrm{f}}$: the probability that the error is reduced is maximum for $K^{\mathrm{o}} = 1$, i.e. with the assimilation of the new observation and the rejection of the first one. Again, the observation with the smaller scale factor is preferred and the other is rejected in this "wild tail" regime $\mu \leq 1$.

2.3. Properties of the "Lévy-estimator" solution

We now examine the fundamental properties of the optimal weight K^{o} given by (12) which holds when the distributions of errors are pure Lévy laws as well as when they only exhibit a power law tail controlling the large variations. Fig. 1 shows the influence of the tail exponent μ on the optimal weight K^{o} as a function of the error ratio λ of the two measurements given by (13). As μ approaches 1, K^{o} crosses over very sharply from 1 to 0 when λ goes through 1, recovering the regime $\mu \leq 1$ given by (15). For larger μ's, the transition of K^{o} from 1 to 0 is smoother as λ varies.

The result (12) for $\mu = 2$ holds not only for the power law distribution itself with $\lambda = (C^{\mathrm{o}})^{1/2}/(C^{\mathrm{f}})^{1/2}$ but also for the Gaussian law distribution with the weight

$$K^{\mathrm{G}} = \frac{1}{1+(\lambda^{\mathrm{G}})^2}, \tag{21}$$

where the superscript $\{\cdot\}^{\mathrm{G}}$ corresponds to the Gaussian, and

$$\lambda^{\mathrm{G}} \equiv \frac{\sigma^{\mathrm{o}}}{\sigma^{\mathrm{f}}} \tag{22}$$

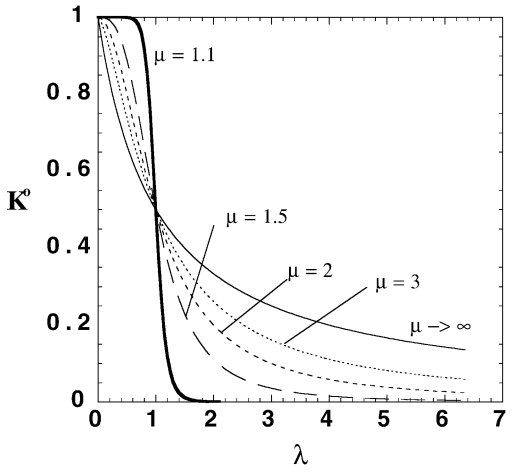

Fig. 1. Dependence of the weight K^{o} given by (12) to the second measurement as a function of the relative amplitude λ (Eq. (13)) of the noise of the two measurements: a small (resp. large) value of λ corresponds to a small (resp. large) error on the second measurement relative to the first one. The different curves correspond to different tail exponents μ.

is the ratio of the characteristic error size according to the Gaussian law. Such K^{G} minimizes the variance $\hat{\sigma}^2$ given by (8). This is natural since the stable Lévy law with $\mu = 2$ is nothing but the Gaussian law with the exact correspondence $C^{\mathrm{f,o}} = (\sigma^{\mathrm{f,o}})^2$ (see Appendix A: there should be no confusion between the exponents μ defining the characteristic functions of stable laws and the exponents μ of arbitrary power laws). Thus, the curve for $\mu = 2$ in Fig. 1 also applies for the Gaussian law with $K^{\mathrm{o}} = K^{\mathrm{G}}$ and $\lambda = \lambda^{\mathrm{G}}$. The result (12) reflects the impact of the relative uncertainties in x^{f} and x^{o} that are quantified by a parameter depending on the ratio of characteristic error size $\lambda^{\mu/(\mu-1)} = (C^{\mathrm{o}}/C^{\mathrm{f}})^{1/(\mu-1)}$.

The weight K^{o} also represents the normalized increment $(\hat{x} - x^{\mathrm{f}})$ added to the initial difference $(x^{\mathrm{o}} - x^{\mathrm{f}})$ as seen from (3)

$$K^{\mathrm{o}} = \frac{\hat{x} - x^{\mathrm{f}}}{x^{\mathrm{o}} - x^{\mathrm{f}}}. \tag{23}$$

Fig. 1 therefore can be interpreted as showing the normalized increment depending on the tail exponent μ, with the extreme cases $\hat{x} = x^{\mathrm{o}}$ at $K^{\mathrm{o}} = 1$ and $\hat{x} = x^{\mathrm{f}}$ at $K^{\mathrm{o}} = 0$. At $\lambda = 1$, where x^{f} and x^{o} has the same uncertainty in terms of scale factor $C^{\mathrm{f}} = C^{\mathrm{o}}$ (13), $K^{\mathrm{o}} = 0.5$ puts $\hat{x}$ at the exact center point between x^{f} and x^{o} for any μ. For $\lambda < 1$ (resp. $\lambda > 1$), where x^{o} with scale factor C^{o} is more (resp. less) accurate than x^{f} with scale factor C^{f}, the smaller the exponent μ is, the closer $\hat{x}$ is to the more accurate sample x^{o} (resp. x^{f}). The estimation by the weight K^{o} for the heavier tail distributions with $\mu < 2$ therefore favors the accurate sample more strongly than in the least-variance case. Interestingly, this situation is reversed for power law tails with exponents $\mu > 2$, i.e. the weight favors the accurate sample less strongly than in the Gaussian case. This situation applies in particular to exponential distributions that are formally obtained as the limit $\mu \to \infty$.

2.4. Quality of improvements: Lévy versus Gaussian estimators

2.4.1. Case $\mu > 2$

Let us investigate the pros and cons of the solution (12) for K^{o} as the optimal weight for power law tails in contrast to its Gaussian counterpart (21) giving K^{G}. In this goal, we propose a specific example using the Student's distribution with μ degrees of freedom, whose density function [13]

$$P_\mu(\omega) = \frac{\Gamma(\frac{1}{2}(\mu+1))}{\sqrt{\mu\pi}\,\Gamma(\frac{1}{2}\mu)} \frac{1/s}{[1 + (\omega/s\sqrt{\mu})^2]^{(1+\mu)/2}}, \tag{24}$$

is defined for $-\infty < \omega < +\infty$. The Student's distribution $P_\mu(\omega)$ has a bell-like shape like the Gaussian (and actually tends to the Gaussian in the limit $\mu \to \infty$). It is, however, a power law like (5) for large $|\omega|$ with a tail exponent equal to the number μ of degrees of freedom defining the Student's distribution, with a scale factor

$$C_\mu(s) = \frac{\Gamma(\frac{1}{2}(\mu+1))}{\sqrt{\mu\pi}\,\Gamma(\frac{1}{2}\mu)} \mu^{(1+\mu)/2} s^\mu. \tag{25}$$

The parameter s represents the typical width of the Student's distribution. The variance exists only for $\mu > 2$ and is given by

$$\mathrm{Var} \equiv \sigma^2 = \frac{\mu}{\mu - 2} s^2. \tag{26}$$

We assume that the forecast (resp. observation) sample x^{f} (resp. x^{o}) has an error ω^{f} (resp. ω^{o}) distributed according to the Student's distribution (24) with typical width s^{f} (resp. s^{o}), but with the same exponent μ. The Lévy

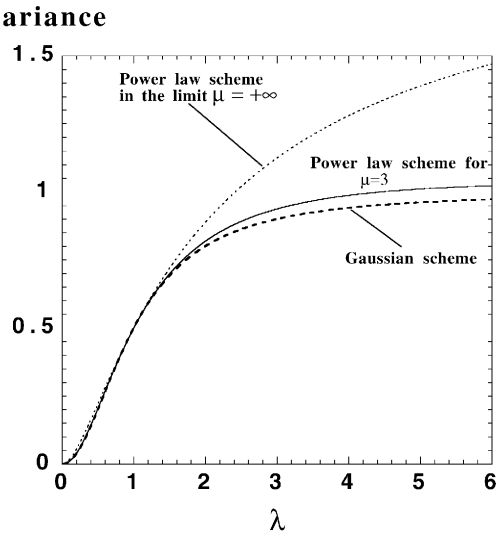

Fig. 2. Dependence of the variance $\mathrm{Var}^{\mathrm{L}}_{\hat{x}}$ and $\mathrm{Var}^{\mathrm{G}}_{\hat{x}}$ of the total error obtained using, respectively, the Lévy and the Gaussian weights, as a function of $\lambda = s^{\mathrm{o}}/s^{\mathrm{f}}$ equal to the ratio of the typical widths of the Student's distributions for the two measurements.

weight K^{L} given by (12) and the standard Gaussian weight K^{G} given by (21) are represented by the same error ratio

$$\lambda^{\mathrm{G}} = \lambda^{\mathrm{L}} = \frac{s^{\mathrm{o}}}{s^{\mathrm{f}}} = \lambda. \tag{27}$$

It is worth recalling that K^{o}, which we denote here K^{L}, given by (12) is obtained so as to minimize the scale factor $C_{\hat{x}}$ given by (7), while K^{G} given by (21) minimizes the variance $\mathrm{Var}_{\hat{x}}$ expressed by (8). The impact of the difference between these two weights can be quantified in several ways for $\mu > 2$ where the variance exists.

One measure is the corresponding variance $\mathrm{Var}_{\hat{x}} = (1-K^{\mathrm{o}})^2(\sigma^{\mathrm{f}})^2 + (K^{\mathrm{o}})^2(\sigma^{\mathrm{o}})^2$ of the total error $(1-K^{\mathrm{o}})\omega^{\mathrm{f}} + K^{\mathrm{o}}\omega^{\mathrm{o}}$ given by

$$\mathrm{Var}^{\mathrm{L}}_{\hat{x}} = \frac{(1+\lambda^{2/(\mu-1)})\lambda^2}{(1+\lambda^{\mu/(\mu-1)})^2}(\sigma^{\mathrm{f}})^2, \tag{28}$$

$$\mathrm{Var}^{\mathrm{G}}_{\hat{x}} = \frac{\lambda^2}{1+\lambda^2}(\sigma^{\mathrm{f}})^2, \tag{29}$$

where $\mathrm{Var}^{\mathrm{L}}_{\hat{x}}$ (resp. $\mathrm{Var}^{\mathrm{G}}_{\hat{x}}$) is the variance obtained by using the solution K^{L} (resp. K^{G}). Fig. 2 shows $\mathrm{Var}^{\mathrm{L}}_{\hat{x}}$ and $\mathrm{Var}^{\mathrm{G}}_{\hat{x}}$ as a function of λ for $\mu = 3$: by construction, we verify that the variance of the total error is less with K^{G} than with K^{o}. This is expected since, by construction, K^{G} minimizes the variance. However, the difference is small, less than 10%. Anyway, this measure would then suggest that the Gaussian filter is better.

However, for power law distributions of errors, the variance is well-known to be a rather poor representation of the variability, especially in the tail. It is thus interesting to compare the scale factors $C^{\mathrm{L}}_{\hat{x}}$ and $C^{\mathrm{G}}_{\hat{x}}$ obtained in the two schemes since they quantify the total weight of the power law tails. We determine the scale factors for the Lévy and Gaussian weights as another measure of the goodness of the filtering method:

$$C^{\mathrm{L}}_{\hat{x}} = \frac{\lambda^{\mu}}{(1+\lambda^{\mu/(\mu-1)})^{\mu-1}}C^{\mathrm{f}}, \tag{30}$$

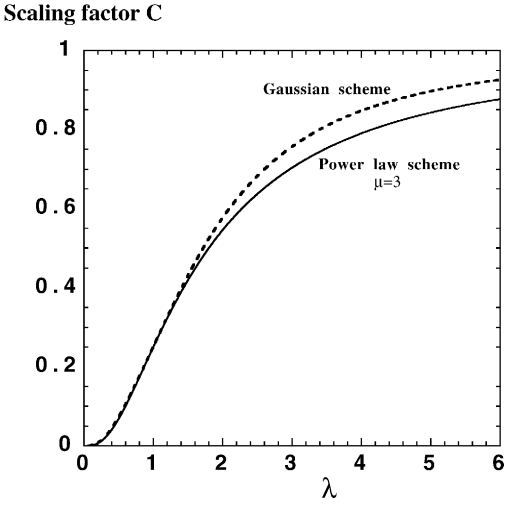

Fig. 3. Dependence of the scale factors $C^{\mathrm{L}}_{\hat{x}}$ and $C^{\mathrm{G}}_{\hat{x}}$ of the total error obtained using, respectively, K^{o} and K^{G} weights, as a function of λ.

$$C^{\mathrm{G}}_{\hat{x}} = \frac{\lambda^{\mu}}{(1+\lambda^2)^{\mu}(1+\lambda^{\mu})^{-1}} C^{\mathrm{f}}. \tag{31}$$

$C^{\mathrm{L}}_{\hat{x}}$ (resp. $C^{\mathrm{G}}_{\hat{x}}$) is obtained by putting K^{L} (resp. K^{G}) in expression (7). Fig. 3 shows the scale factors $C^{\mathrm{L}}_{\hat{x}}$ and $C^{\mathrm{G}}_{\hat{x}}$ as a function of λ for $\mu = 3$: the weight K^{L} is now found to be better than the usual Gaussian weight K^{G}, since a smaller scale factor implies smaller probabilities for large fluctuations. The improvement is however not very large, typically of the order of or less than 10%, i.e. of the same order as the difference between the variances (but in reverse ranking). These relatively small differences between the Gaussian and Lévy filtering procedures become enormous for the case $\mu < 2$ discussed next.

The comparison between Figs. 2 and 3 shows that one cannot achieve simultaneously the minimization of the variance of the error and the minimization of the weight of the tail of large deviations of the error: either one or the other can be optimized.

2.4.2. *Case $\mu < 2$*

The situation is dramatically different when $\mu < 2$ for which the variance is not mathematically defined. In this case, an empirical determination of the variance is very unstable and absolutely unreliable. The standard Gaussian weight K^{G} is completely useless. In contrast, the Lévy weight K^{L} gives a simple and clear-cut recipe that allows one to optimize large fluctuations in the weighting procedure.

Let us illustrate this result by the following numerical experiments using the Cauchy distribution for the errors:

$$P_C(\omega) = \frac{1/s}{(\omega/s)^2 + \pi^2} \tag{32}$$

with typical width s. The Cauchy distribution (32) is one of the stable Lévy distribution and possesses a power law tail with exponent $\mu = 1$ and a scale factor $C = s$. Let us assume that the first (resp. second) sample x^{f} (resp. x^{o}) has an error ω^{f} (resp. ω^{o}) distributed according to the Cauchy distribution (32) with typical width s^{f} (resp. s^{o}). Then, the resulting distribution of the errors on $\hat{x}$ is of the same form (32) with the same exponent $\mu = 1$, while the scale factor is given by

$$C_{\hat{x}} = (1 - K^{\mathrm{o}})s^{\mathrm{f}} + K^{\mathrm{o}}s^{\mathrm{o}}. \tag{33}$$

As we have found above in (15), the weight K^{o} that minimizes $C_{\hat{x}}$ is

$$K^{\mathrm{o}} = \begin{cases} 0 & \text{for } s^{\mathrm{o}} > s^{\mathrm{f}}, \\ 1 & \text{for } s^{\mathrm{o}} < s^{\mathrm{f}}, \end{cases} \tag{34}$$

since the Cauchy distribution is on the borderline $\mu = 1$. This can be verified straightforwardly as a result of the linear dependence of $C_{\hat{x}}$ on K^{o} in (33) for which the optimization always selects one of the boundaries.

Consider the case where $s^{\mathrm{f}} = 1$ and $s^{\mathrm{o}} = 2$. The Lévy estimator imposes to choose $K^{\mathrm{o}} = 0$, i.e. to reject the information provided by the second sample x^{o}. Let us now compare this recipe with the result obtained by applying the standard Gaussian weight K^{G} on the data generated by using the two Cauchy laws with $s^{\mathrm{f}} = 1$ and $s^{\mathrm{o}} = 2$. Specifically, we generated two sets of 1000 random numbers ω^{f} and ω^{o}, distributed according to the Cauchy law (32) with $s^{\mathrm{f}} = 1$ and $s^{\mathrm{o}} = 2$. From each of these 1000 numbers, we can estimate numerically the variance and find $(\sigma^{\mathrm{f}})^2 = 3.36 \times 10^5$ and $(\sigma^{\mathrm{o}})^2 = 4.07 \times 10^5$. The Gaussian estimator (21) then recommends the value $K^{\mathrm{G}} = [1 + (\sigma^{\mathrm{o}}/\sigma^{\mathrm{f}})^2]^{-1} \approx 0.45$ which is very different from $K^{\mathrm{o}} = 0$ given by (34). We should stress that the estimations of the variances $(\sigma^{\mathrm{f}})^2$ and $(\sigma^{\mathrm{o}})^2$ are highly unreliable because they can change by orders of magnitude from one sample to another. The reason is, as we have said, that the variance is mathematically infinite in this case, and therefore any estimation of it is bound to be dominated by the few largest random numbers that occur by chance in the series.

Two lessons are thus to be learned from these numerical simulations: (i) estimating the variance for distributions with exponent $\mu < 2$ leads to very unstable results; (ii) the resulting recommendation of the standard Gaussian weight K^{G} can be very wrong.

3. Multivariate estimator

3.1. Definition of the model

When the state variables are multi-dimensional, their errors may be mutually dependent. If the errors are distributed according to the power or Lévy laws, we need to transform the coordinate of errors to express them as linear sums of independent noises in order to be able use the rules stated above in points (1)–(4).

We consider a problem of estimating a multi-dimensional state vector $\mathbf{x} \in \mathbb{R}^N$ using two samples $\mathbf{x}^{\mathrm{f}} \in \mathbb{R}^N$ and $\mathbf{x}^{\mathrm{o}} \in \mathbb{R}^N$. Both forecast and observations are made for all state variables. As in the case for the univariate estimation, we assume that the estimate $\hat{\mathbf{x}}$ of $\mathbf{x}$ can be expressed as a linear combination of $\mathbf{x}^{\mathrm{f}}$ and $\mathbf{x}^{\mathrm{o}}$. Requiring the unbiased condition leads to one unknown weight matrix in the estimation

$$\hat{\mathbf{x}} = (\mathbf{I} - \mathbf{K}^{\mathrm{o}})\mathbf{x}^{\mathrm{f}} + \mathbf{K}^{\mathrm{o}}\mathbf{x}^{\mathrm{o}}, \tag{35}$$

where $\mathbf{I}$ is an identity matrix and $\mathbf{K}^{\mathrm{o}} \in \mathbb{R}^{N\times N}$ the weight matrix for the observation sample $\mathbf{x}^{\mathrm{o}}$. Our goal is to determine the optimal $\mathbf{K}^{\mathrm{o}}$ which gives the least uncertainty in $\hat{\mathbf{x}}$. We use the notation $\mathbf{K}$ for the weight matrix in connection to the standard Kalman(–Gaussian) gain matrix of sequential estimation [12].

We assume that the sample error vectors are linear sums of N independent μ-variables with symmetric distributions

$$\boldsymbol{\epsilon}^{\mathrm{f,o}} = \mathbf{x}^{\mathrm{f,o}} - \mathbf{x} = \mathbf{G}^{\mathrm{f,o}}\boldsymbol{\omega}^{\mathrm{f,o}}, \tag{36}$$

where the probability distribution of each $\omega_l^{\mathrm{f,o}}$ is associated with the same exponent μ and usually different individual scale factors $C_l^{\mathrm{f,o}}$. The assumption that the distributions of $\boldsymbol{\omega}^{\mathrm{f,o}}$'s are symmetric implies that the same scale factors $C_l^{\mathrm{f,o}}$ characterize the tail of large positive and negative realizations. The transformation between independent $\boldsymbol{\omega}^{\mathrm{f,o}}$ to mutually dependent $\boldsymbol{\epsilon}^{\mathrm{f,o}}$ is provided by the matrix $\mathbf{G}^{\mathrm{f,o}} \in \mathbb{R}^{N\times N}$. We shall use this decomposition scheme

repeatedly in the sequel as it allows us to treat in a simple way the interplay between the power law distributions and the dependence between variables.

In the standard linear least-variance estimation theory, one calculates the covariance matrix $\hat{\mathbf{P}} \equiv \langle \boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}} \rangle$ of the error $\boldsymbol{\epsilon} = \hat{\mathbf{x}} - \mathbf{x}$ and minimizes the expectation of the distance between $\hat{\mathbf{x}}$ and $\mathbf{x}$ in the mean-square sense, i.e., one minimizes the trace of the covariance matrix $\hat{\mathbf{P}}$. The covariance calculation is an essential step to guarantee that all error components are suitably accounted for. We thus propose our key idea to generalize the covariance matrix in the regime $\mu < 2$, where it does not exist by using the concept of the tail-covariance defined as the matrix of scale factors of the distribution of all products $\epsilon_i^{\mathrm{f,o}} \epsilon_j^{\mathrm{f,o}}$. We prefer this approach to the so-called "covariation" [21] as it is more intuitive and also presents nicer properties, in particular the tail-covariance matrix remains symmetric. The intuitive meaning of the covariation is less transparent than for the tail-covariance, which explicitly measures the correlations between large events only, while the covariation picks up contributions from the core of the distributions. Let us mention that a simplified version of the tail-covariance without signs (see below) has been used in the context of portfolio theory [3].

Consider the sample error vectors $\boldsymbol{\epsilon}^{\mathrm{f,o}}$. We thus study the product $\epsilon_i^{\mathrm{f,o}} \epsilon_j^{\mathrm{f,o}}$, whose probability distribution constitutes the natural generalization of the covariance as already pointed out. Using properties (3) and (4) above, one finds the following result, expressed symbolically and without the superscripts for simplicity:

$$\epsilon_i \epsilon_j \simeq \sum_{l}^{N} |G_{il}||G_{jl}|[\tfrac{1}{2}\mu\text{-variable}] + \sum_{l'}^{N} \sum_{l'' \neq l'}^{N} |G_{il'}||G_{jl''}|[\mu\text{-variable}], \tag{37}$$

where the symbol "μ-variable" defines a random variable distributed with a density distribution with a power law tail of exponent μ. Expression (37) means that the tail of the distribution of the product $\epsilon_i \epsilon_j$ is *dominated by the first term* which has the smallest exponent $\frac{1}{2}\mu$ and thus heaviest tail, and will directly be sensitive to the product $|G_{il}||G_{jl}|$ for all the independent errors ω_l. More precisely, an analysis of the cumulative distribution of the products $\epsilon_i \epsilon_j$ will give an asymptotic slope of $-\frac{1}{2}\mu$ in a log–log plot and a scale factor C_l associated with ω_l proportional to $|G_{il}|^{\mu}|G_{jl}|^{\mu}$.

We use this set of scale factors associated with the distributions of $\epsilon_i \epsilon_j$ in order to define the tail-covariance matrix $\mathbf{B} \in \mathbb{R}^{N \times N}$ with the following guidelines. It is natural that the tail-covariance matrix should contain the information on the tail of the products $\epsilon_i \epsilon_j$ with distribution (37). A bonafide generalization of the covariance matrix requires two additional conditions. It should be sensitive to the sign of the dependence between the variables, i.e., if ϵ_i increases (resp. decreases) on average conditioned on the increase of ϵ_j, the dependence is positive (resp. negative), generalizing the existence of positive and negative correlations. In addition, a suitable definition of the tail-covariance matrix should be such that it recovers the standard covariance matrix for the value $\mu = 2$ corresponding to the special case where the stable laws reduce to the Gaussian distribution, as briefly recalled in Appendix A. These considerations lead to the following unique specification for the tail-covariance matrix by diagonalization:

$$\mathbf{B} = \mathbf{G}^{[\mu/2]} \mathbf{C} \mathbf{G}^{\mathrm{T}[\mu/2]}, \tag{38}$$

where $\mathbf{C} \in \mathbb{R}^{N \times N}$ is a diagonal matrix associated with $\boldsymbol{\omega}$ and $\mathbf{G}^{[\mu/2]} \in \mathbb{R}^{N \times N}$ is the corresponding eigenvector matrix. The operator $\{\cdot\}^{[\beta]}$ means that each element of matrix or vector is defined by

$$\mathbf{G}_{ij}^{[\beta]} = \mathrm{sign}(G_{ij})|G_{ij}|^{\beta}, \tag{39}$$

i.e., the absolute value of each element is raised to the power β and then multiplied by its sign. This operation can be applied to scalars as well. Without the sign operator in (39), the tail-covariance matrix $\mathbf{B}$ would be just the matrix of scale factors of all the products $\epsilon_i \epsilon_j$. The introduction of the sign function is an essential additional ingredient introduced to account for the sign of the dependence between the variables.

For $\mu = 2$, $\mathbf{G}_{ij}^{[\mu/2]} = \mathrm{sign}(G_{ij})|G_{ij}| = G_{ij}$, and we check that the tail-covariance is exactly the same as the error covariance

$$\mathbf{B} = \mathbf{P} \equiv \langle \boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}} \rangle. \tag{40}$$

This correspondence may appear paradoxical if one interprets the case $\mu = 2$ as corresponding to a power law, which has infinite variance. As we already discussed for the unidimensional case, the technical reason for this comes from the form (9) of the characteristic function $\hat{L}_\mu(k)$ of a distribution with a power law tail, as given in [7] (Problem 15 of Section 12) or of a symmetric Lévy laws. We stress again the important point that for μ strictly less than 2, the inverse Fourier transform of $\hat{L}_\mu(k)$ defined by (9) gives a power law tail, while for $\mu = 2$, it gives a Gaussian law. The continuity between the expressions (38) and (40) can thus be traced back to that of $\hat{L}_\mu(k)$ as a function of μ at $\mu = 2$.

The transformation of the scale factors between the mutually dependent errors $\boldsymbol{\epsilon}$ and independent errors $\boldsymbol{\omega}$ as in (38) can be performed in both directions, i.e., not only from right- to left-hand side to compute $\mathbf{B}$ when $\mathbf{G}$ and $\mathbf{C}$ are known, but also from left- to right-hand side to obtain $\mathbf{C}$ (and $\mathbf{G}$) by diagonalization of $\mathbf{B}$.

3.2. Solution

Our aim is to obtain the optimal weight $\mathbf{K}^{\mathrm{o}}$ in (35) for the best estimate $\hat{\mathbf{x}}$. In this goal, we form the set of products $\epsilon_i \epsilon_j$, where $\boldsymbol{\epsilon} = \hat{\mathbf{x}} - \mathbf{x}$ and study their probability distribution. As in (37) and (38), we retain only the term decaying as a power law with an exponent $\frac{1}{2}\mu$ and get the following tail-covariance matrix of $\hat{\mathbf{x}} - \mathbf{x}$ for an arbitrary weight matrix $\mathbf{K}^{\mathrm{o}}$:

$$\hat{\mathbf{B}} = (\mathbf{G}^{\mathrm{f}} - \mathbf{K}^{\mathrm{o}}\mathbf{G}^{\mathrm{f}})^{[\mu/2]}\mathbf{C}^{\mathrm{f}}(\mathbf{G}^{\mathrm{f}} - \mathbf{K}^{\mathrm{o}}\mathbf{G}^{\mathrm{f}})^{\mathrm{T}[\mu/2]} + (\mathbf{K}^{\mathrm{o}}\mathbf{G}^{\mathrm{o}})^{[\mu/2]}\mathbf{C}^{\mathrm{o}}(\mathbf{K}^{\mathrm{o}}\mathbf{G}^{\mathrm{o}})^{\mathrm{T}[\mu/2]}. \tag{41}$$

Here, we assume that errors $\boldsymbol{\epsilon}^{\mathrm{f}}$ and $\boldsymbol{\epsilon}^{\mathrm{o}}$ are mutually independent.

In the univariate case, the optimization process is unique and corresponds to minimizing $\hat{C}$ with respect to K^{o} as in (10). In the multivariate case, however, the optimization may be defined in several ways. For example, as recalled above, the standard linear least-variance estimation theory attempts to minimize the expectation distance between $\hat{\mathbf{x}}$ and $\mathbf{x}$ in the mean-square sense, i.e., to minimize trace $\hat{\mathbf{P}}$ given by (40). For $\mu < 2$, where the covariance $\hat{\mathbf{P}}$ does not exist, we propose to minimize the "average" scale factor, i.e., trace $\hat{\mathbf{B}}$. Such an optimization implies that the uncertainty in $\hat{\mathbf{x}}$ is globally the smallest (see Section 2.2 for the univariate case).

Since the expression (41) is not smooth in K^{o} due to the presence of the absolute values, some care must be taken in the minimization. The non-smooth character of (41) makes the differentiation approach to the minimization more cumbersome, as one must keep track of the discontinuities. The optimal $\mathbf{K}^{\mathrm{o}}$ is obtained by solving

$$\frac{\partial}{\partial \mathbf{K}^{\mathrm{o}}} \mathrm{trace}\, \hat{\mathbf{B}} = 0, \tag{42}$$

where

$$\begin{aligned}
\hat{B}_{ii} = &\sum_{p=1}^{N} \left(G_{ip}^{\mathrm{f}} - \sum_{m=1}^{N} K_{im}^{\mathrm{o}} G_{mp}^{\mathrm{f}} \right)^{[\mu/2]} C_p^{\mathrm{f}} \left(G_{ip}^{\mathrm{f}} - \sum_{m=1}^{N} K_{im}^{\mathrm{o}} G_{mp}^{\mathrm{f}} \right)^{[\mu/2]} \\
&+ \sum_{p=1}^{N} \left(\sum_{m=1}^{N} K_{im}^{\mathrm{o}} G_{mp}^{\mathrm{o}} \right)^{[\mu/2]} C_p^{\mathrm{o}} \left(\sum_{m=1}^{N} K_{im}^{\mathrm{o}} G_{mp}^{\mathrm{o}} \right)^{[\mu/2]} \\
= &\sum_{p=1}^{N} \left| G_{ip}^{\mathrm{f}} - \sum_{m=1}^{N} K_{im}^{\mathrm{o}} G_{mp}^{\mathrm{f}} \right|^{\mu} C_p^{\mathrm{f}} + \sum_{p=1}^{N} \left| \sum_{m=1}^{N} K_{im}^{\mathrm{o}} G_{mp}^{\mathrm{o}} \right|^{\mu} C_p^{\mathrm{o}}.
\end{aligned} \tag{43}$$

Taking the derivative with respect to K^{o}_{ij}, we obtain

$$\frac{\partial}{\partial K^{\mathrm{o}}_{ij}} \operatorname{trace} \hat{\mathbf{B}} = \frac{\partial}{\partial K^{\mathrm{o}}_{ij}} \hat{B}_{ii} = \mu \sum_{p=1}^{N} \left[-\left(G^{\mathrm{f}}_{ip} - \sum_{m=1}^{N} K^{\mathrm{o}}_{im} G^{\mathrm{f}}_{mp} \right)^{[\mu-1]} G^{\mathrm{f}}_{jp} C^{\mathrm{f}}_{p} + \left(\sum_{m=1}^{N} K^{\mathrm{o}}_{im} G^{\mathrm{o}}_{mp} \right)^{[\mu-1]} G^{\mathrm{o}}_{jp} C^{\mathrm{o}}_{p} \right] = 0, \tag{44}$$

where we have used the fact that $\mathrm{d}|x|^{\mu}/\mathrm{d}x = \mu \operatorname{sign}(x)|x|^{\mu-1} = \mu x^{[\mu-1]}$, with our definition (39). Note that the same N coefficients K^{o}_{im} with $m = 1, \ldots, N$ and only them occur in the N equations $(\partial/\partial K^{\mathrm{o}}_{ij}) \operatorname{trace} \hat{\mathbf{B}}$ obtained by fixing i and varying j from 1 to N. Finding the optimal $\mathbf{K}^{\mathrm{o}}$ thus involves solving N independent systems of equations, one for each individual diagonal term $\hat{B}_{ii}$, where each system consists of the N nonlinear equations $(\partial/\partial K^{\mathrm{o}}_{ij})\hat{B}_{ii} = 0$ for the N unknowns K^{o}_{ij} at i fixed.

Optimality of $\mathbf{K}^{\mathrm{o}}$ is ensured by the positive definiteness of the Hessian matrix for $\hat{B}_{ii}$ with respect to K^{o}_{ij}, for each of the N independent systems (each defined as we have shown by a fixed i) of N equations (obtained by varying j at fixed i). While for the univariate case, this amounts to ensure the validity of only one equation (11), in the multivariate case we need to study the positive definiteness of the Hessian

$$\frac{\partial^2}{\partial(K^{\mathrm{o}}_{ij})\partial(K^{\mathrm{o}}_{iq})} \hat{B}_{ii} = \mu(\mu-1) \left[\sum_{p=1}^{N} (G^{\mathrm{f}}_{jp})(G^{\mathrm{f}}_{qp}) \left| G^{\mathrm{f}}_{ip} - \sum_{m=1}^{N} K^{\mathrm{o}}_{im} G^{\mathrm{f}}_{mp} \right|^{\mu-2} C^{\mathrm{f}}_{p} + \sum_{p=1}^{N} (G^{\mathrm{o}}_{jp})(G^{\mathrm{o}}_{qp}) \left| \sum_{m=1}^{N} K^{\mathrm{o}}_{im} G^{\mathrm{o}}_{mp} \right|^{\mu-2} C^{\mathrm{o}}_{p} \right], \tag{45}$$

expressed at the values of $\mathbf{K}^{\mathrm{o}}$ solving (44). If this matrix is positive definite, then K^{o}_{ij} minimizes $\hat{B}_{ii}$ and is therefore optimal. There are two issues associated with this formulation for the optimal $\mathbf{K}^{\mathrm{o}}$: (1) the possible existence of singular terms for $\mu < 2$; and (2) the solvability of the nonlinear system. We address them in Appendix B and show that the matrix is indeed positive definite and that there is only one solution.

3.3. Special cases

3.3.1. Case $\mu = 2$

For $\mu = 2$, the Lévy estimator is the same as the Gaussian one which minimizes $\operatorname{trace} \hat{\mathbf{P}} = \operatorname{trace} \hat{\mathbf{B}}$ as given by (40). In this case, the system becomes perfectly linear with $\frac{1}{2}\mu = \mu - 1 = 1$ and leads to the following analytical solution for the optimal weight given in matrix form

$$\mathbf{K}^{\mathrm{G}} = \mathbf{B}^{\mathrm{f}}(\mathbf{B}^{\mathrm{f}} + \mathbf{B}^{\mathrm{o}})^{-1}, \tag{46}$$

which results in the optimal estimates of the state variable and corresponding covariance matrix

$$\hat{\mathbf{x}}^{\mathrm{G}} = \mathbf{B}^{\mathrm{o}}(\mathbf{B}^{\mathrm{f}} + \mathbf{B}^{\mathrm{o}})^{-1}\mathbf{x}^{\mathrm{o}} + \mathbf{B}^{\mathrm{f}}(\mathbf{B}^{\mathrm{f}} + \mathbf{B}^{\mathrm{o}})^{-1}\mathbf{x}^{\mathrm{f}}, \tag{47}$$

$$\hat{\mathbf{B}}^{\mathrm{G}} = \mathbf{B}^{\mathrm{o}}(\mathbf{B}^{\mathrm{f}} + \mathbf{B}^{\mathrm{o}})^{-1}\mathbf{B}^{\mathrm{f}}. \tag{48}$$

3.3.2. Independent noise

When the errors $\boldsymbol{\epsilon}^{\mathrm{f,o}}$ are independent, i.e., $\mathbf{G}^{\mathrm{f,o}} = \mathbf{I}$, the trace and diagonal components (43) for the average scale factor can be simplified into

$$\operatorname{trace} \hat{\mathbf{B}} = \sum_{i=1}^{N} \sum_{p=1}^{N} [(1 - K^{\mathrm{o}}_{ip})^{\mu} C^{\mathrm{f}}_{p} + (K^{\mathrm{o}}_{ip})^{\mu} C^{\mathrm{o}}_{p}], \tag{49}$$

where we have omitted the absolute values and the sign functions as we look for values of the Kalman weights between 0 and 1. Therefore, the problem can be reduced to the univariate estimate for each element x_i independently as defined in Eq. (7). Eqs. (10) and (11) are thus replaced by their exact equivalent derived from the two conditions (42) and (45). The solution K_{ii}^{o} for each i is given by (12), where $K_{ii}^{\mathrm{o}} = K^{\mathrm{o}}$ is obtained by setting $\lambda = (C_i^{\mathrm{o}})^{1/\mu}/(C_i^{\mathrm{f}})^{1/\mu}$ (see (13)). Due to the hypothesis of noise independence, the non-diagonal coefficients K_{ij}^{o} are all zero for $i \neq j$.

4. The KL filter

4.1. Problem

We are now in a position to construct a sequential data assimilation methodology when the noises are distributed according to the power or Lévy law. The standard assimilation problem is formulated as follows. We consider a linear discrete stochastic dynamical system of state variables $\mathbf{x} \in \mathbb{R}^N$

$$\mathbf{x}_k^{\mathrm{t}} = \mathbf{M}_{k,k-1}\mathbf{x}_{k-1}^{\mathrm{t}} + \boldsymbol{\eta}_{k-1}^{\mathrm{t}}, \tag{50}$$

where superscript $\{\cdot\}^{\mathrm{t}}$ denotes the true state and $\boldsymbol{\eta}_{k-1}^{\mathrm{t}}$ is the dynamical noise. The index k corresponds to the time sequence when the observations $\mathbf{y}_k^{\mathrm{o}} \in \mathbb{R}^{L_k}$ are taken as

$$\mathbf{y}_k^{\mathrm{o}} = \mathbf{H}_k\mathbf{x}_k^{\mathrm{t}} + \boldsymbol{\epsilon}_k^{\mathrm{o}}, \tag{51}$$

where $\boldsymbol{\epsilon}_k^{\mathrm{o}}$ is the observational noise and $\mathbf{H}_k \in \mathbb{R}^{L_k \times N}$ the linear observation function which can vary at each time step k. These observations $\mathbf{y}_k^{\mathrm{o}}$ are assumed to be linear functions of the state variable $\mathbf{x}_k^{\mathrm{t}}$ of the system with an additive noise.

The estimation methodology developed in the previous sections is now extended to perform a filtering in order to estimate $\mathbf{x}^{\mathrm{t}}$ sequentially, by assimilating $\mathbf{y}_k^{\mathrm{o}}$ into the deterministic forecast $\mathbf{x}_k^{\mathrm{f}}$. Here, we assume that $\mathbf{M}_{k,k-1}$ and $\mathbf{H}_k$ are known. The assimilation cycle k is defined over a time interval $[k-1, k]$ between the two adjacent observation events. It consists of the following two steps:

1. A deterministic forecast $\mathbf{x}_k^{\mathrm{f}}$ of $\mathbf{x}_k^{\mathrm{t}}$ from a given initial condition $\mathbf{x}_{k-1}^{\mathrm{a}}$ as the best estimate of $\mathbf{x}_{k-1}^{\mathrm{t}}$ based on the model

$$\mathbf{x}_k^{\mathrm{f}} = \mathbf{M}_{k,k-1}\mathbf{x}_{k-1}^{\mathrm{a}}. \tag{52}$$

   The forecast is based on the analysis performed at the previous time step.
2. This forecast is then used to construct the new analysis $\mathbf{x}_k^{\mathrm{a}}$ which is mixed with the assimilated observation. This leads to the probabilistic analysis $\mathbf{x}_k^{\mathrm{a}}$ of $\mathbf{x}_k^{\mathrm{t}}$ obtained as the weighted average of $\mathbf{x}_k^{\mathrm{f}}$ and $\mathbf{y}_k^{\mathrm{o}}$ under the unbias assumption at time k:

$$\mathbf{x}_k^{\mathrm{a}} = \mathbf{x}_k^{\mathrm{f}} + \mathbf{K}_k(\mathbf{y}_k^{\mathrm{o}} - \mathbf{H}_k\mathbf{x}_k^{\mathrm{f}}) = (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{x}_k^{\mathrm{f}} + \mathbf{K}_k\mathbf{y}_k^{\mathrm{o}}. \tag{53}$$

Accordingly, the errors associated with $\mathbf{x}_k^{\mathrm{f}}$ and $\mathbf{x}_k^{\mathrm{a}}$ are auto-regressive processes.

$$\mathbf{x}_k^{\mathrm{f}} - \mathbf{x}_k^{\mathrm{t}} = \mathbf{M}_{k,k-1}(\mathbf{x}_{k-1}^{\mathrm{a}} - \mathbf{x}_{k-1}^{\mathrm{t}}) - \boldsymbol{\eta}_{k-1}^{\mathrm{t}}, \tag{54}$$

$$\mathbf{x}_k^{\mathrm{a}} - \mathbf{x}_k^{\mathrm{t}} = (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)(\mathbf{x}_k^{\mathrm{f}} - \mathbf{x}_k^{\mathrm{t}}) + \mathbf{K}_k\boldsymbol{\epsilon}_k^{\mathrm{o}} = (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{M}_{k,k-1}(\mathbf{x}_{k-1}^{\mathrm{a}} - \mathbf{x}_{k-1}^{\mathrm{t}}) - (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\boldsymbol{\eta}_{k-1}^{\mathrm{t}} + \mathbf{K}_k\boldsymbol{\epsilon}_k^{\mathrm{o}}, \tag{55}$$

and the only unknown to be determined is the so-called "gain matrix" $\mathbf{K}_k$ needed in order to complete the assimilation cycle.

Our goal is therefore to determine the gain matrix $\mathbf{K}_k^{\mathrm{L}}$, where the superscript L refers to the KL filter, which results in the least uncertainty in $\mathbf{x}_k^{\mathrm{a}}$ in each assimilation cycle when the noises are distributed according to the power or Lévy law with the exponent μ. As in (36), we express the sample error vectors as linear sums of N independent μ-variables

$$\boldsymbol{\eta}_k \equiv \mathbf{G}_k^{\eta}\boldsymbol{\omega}_k^{\eta}, \qquad \boldsymbol{\epsilon}_k \equiv \mathbf{G}_k^{\epsilon}\boldsymbol{\omega}_k^{\epsilon}. \tag{56}$$

Because (54) and (55) define linear autoregressive processes with Lévy-stable or power law probability distribution of the noises, the errors in forecast and analysis are also distributed according to the power or Lévy laws with the same exponent μ. Without loss of generality, they can thus be written as

$$\mathbf{x}_k^{\mathrm{f,a}} - \hat{\mathbf{x}}_k \equiv \mathbf{G}_k^{\mathrm{f,a}}\boldsymbol{\omega}_k^{\mathrm{f,a}}, \tag{57}$$

which defines the matrices $\mathbf{G}_k^{\mathrm{f,a}}$ and the vectors $\boldsymbol{\omega}_k^{\mathrm{f,a}}$ of independent Lévy or power law processes. Consequently, all error and noise distributions in the assimilation cycles are characterized by the corresponding tail-covariance matrices

$$\mathbf{B}_k^{\mathrm{f,a}} \equiv (\mathbf{G}_k^{\mathrm{f,a}})^{[\mu/2]}\mathbf{C}_k^{\mathrm{f,a}}(\mathbf{G}_k^{\mathrm{f,a}})^{\mathrm{T}[\mu/2]}, \tag{58}$$

$$\mathbf{B}_k^{\eta,\epsilon} \equiv (\mathbf{G}_k^{\eta,\epsilon})^{[\mu/2]}\mathbf{C}_k^{\eta,\epsilon}(\mathbf{G}_k^{\eta,\epsilon})^{\mathrm{T}[\mu/2]}, \tag{59}$$

where $\mathbf{C} \in \mathbb{R}^{N\times N}$ is a diagonal scale-factor matrix associated with $\boldsymbol{\omega}$ and $\mathbf{G}^{[\mu/2]}$ is the corresponding eigenvector matrix. Given $\mathbf{B}$, the corresponding $\mathbf{G}$ and $\mathbf{C}$ can be obtained by diagonalization.

4.2. Solution

The optimal KL filter $\mathbf{K}_k^{\mathrm{L}}$ which minimizes the global average error is obtained by minimizing the trace of the tail-covariance matrix $\mathbf{B}_k^{\mathrm{a}}$ of the resulting probabilistic analysis $\mathbf{x}_k^{\mathrm{a}}$ of $\mathbf{x}_k^{\mathrm{t}}$. Since the expression for $\mathbf{B}_k^{\mathrm{a}}$ is not smooth in $\mathbf{K}_k^{\mathrm{L}}$ due to the presence of the absolute values, some care must be taken in the minimization as already discussed. The non-smooth character of $\mathbf{B}_k^{\mathrm{a}}$ makes the differentiation approach to the minimization more cumbersome, as one must keep track of the discontinuities. However, we use the approach described in Appendix B to check if the two conditions are simultaneously verified

$$\frac{\partial}{\partial \mathbf{K}_k^{\mathrm{L}}} \operatorname{trace} \mathbf{B}_k^{\mathrm{a}} = 0, \tag{60}$$

and positive-definiteness of the Hessian matrix

$$\frac{\partial^2}{\partial (\mathbf{K}_k^{\mathrm{L}})^2} \operatorname{trace} \mathbf{B}_k^{\mathrm{a}}. \tag{61}$$

We solve the first condition and then examine the second condition. This is achieved by taking the following two steps in each assimilation cycle:

Step 1.

Dynamic forecast: Given a set of initial conditions described by the subscript $\{\cdot\}_{k-1}$ which are known, the forecast is performed deterministically to advance from $k-1$ to k based on (52) and (54)

$$\mathbf{x}_k^{\mathrm{f}} = \mathbf{M}_{k,k-1}\mathbf{x}_{k-1}^{\mathrm{a}} \tag{62}$$

leading to the tail-covariance of the forecasts at time k:

$$\mathbf{B}_k^{\mathrm{f}} = (\mathbf{M}_{k,k-1}\mathbf{G}_{k-1}^{\mathrm{a}})^{[\mu/2]}\mathbf{C}_{k-1}^{\mathrm{a}}(\mathbf{M}_{k,k-1}\mathbf{G}_{k-1}^{\mathrm{a}})^{\mathrm{T}[\mu/2]} + \mathbf{B}_{k-1}^{\eta}, \tag{63}$$

where the definition (38) and (39) is used.

Step 2.

Probabilistic analysis: Given the forecast $\mathbf{x}_k^{\mathrm{f}}$ with $\mathbf{B}_k^{\mathrm{f}}$ from Step 1 along with the observations $\mathbf{y}_k^{\mathrm{o}}$ with tail-covariance $\mathbf{B}_k^{\epsilon}$, the analysis provides the optimal estimate

$$\mathbf{x}_k^{\mathrm{a}} = \mathbf{x}_k^{\mathrm{f}} + \mathbf{K}_k^{\mathrm{L}}(\mathbf{y}_k^{\mathrm{o}} - \mathbf{H}_k\mathbf{x}_k^{\mathrm{f}}) \tag{64}$$

with tail-covariance

$$\mathbf{B}_k^{\mathrm{a}} = (\mathbf{G}_k^{\mathrm{f}} - \mathbf{K}_k^{\mathrm{L}}\mathbf{H}_k\mathbf{G}_k^{\mathrm{f}})^{[\mu/2]}\mathbf{C}_k^{\mathrm{f}}(\mathbf{G}_k^{\mathrm{f}} - \mathbf{K}_k^{\mathrm{L}}\mathbf{H}_k\mathbf{G}_k^{\mathrm{f}})^{\mathrm{T}[\mu/2]} + (\mathbf{K}_k^{\mathrm{L}}\mathbf{G}_k^{\epsilon})^{[\mu/2]}\mathbf{C}_k^{\epsilon}(\mathbf{K}_k^{\mathrm{L}}\mathbf{G}_k^{\epsilon})^{\mathrm{T}[\mu/2]}, \tag{65}$$

where the definition (38) and (39) is used. By letting subscripts represent the matrix elements and dropping the time index k for simplicity, we solve for the optimal filter $\mathbf{K}^{\mathrm{L}}$ so that it satisfies the conditions (60). The diagonal elements of B^{a} can be explicated as follows:

$$B_{ii}^{\mathrm{a}} = \sum_{p=1}^{N}\left|G_{ip}^{\mathrm{f}} - \sum_{m=1}^{L}K_{im}H_{mp}^{\mathbf{G}}\right|^{\mu} C_p^{\mathrm{f}} + \sum_{q=1}^{L}\left|\sum_{m=1}^{L}K_{im}G_{mq}^{\epsilon}\right|^{\mu} C_q^{\epsilon} \tag{66}$$

with

$$\mathbf{H}^{\mathbf{G}} \equiv \mathbf{H}\mathbf{G}^{\mathrm{f}}. \tag{67}$$

Because B_{ii}^{a} does not depend on K_{qj} for $q \neq i$, the first condition for the optimal $\mathbf{K}_k^{\mathrm{L}}$ (60) is

$$\begin{aligned}\frac{\partial}{\partial K_{ij}^{\mathrm{L}}}\operatorname{trace}\mathbf{B}^{\mathrm{a}} &= \frac{\partial}{\partial K_{ij}^{\mathrm{L}}}B_{ii}^{\mathrm{a}} = 0\\ &= \mu\left[-\sum_{p=1}^{N}\left(G_{ip}^{\mathrm{f}} - \sum_{m=1}^{L}K_{im}^{\mathrm{L}}H_{mp}^{\mathbf{G}}\right)^{[\mu-1]} H_{jp}^{\mathbf{G}}C_p^{\mathrm{f}} + \sum_{q=1}^{L}\left(\sum_{m=1}^{L}K_{im}^{\mathrm{L}}G_{mq}^{\epsilon}\right)^{[\mu-1]} G_{jq}^{\epsilon}C_q^{\epsilon}\right] = 0,\end{aligned} \tag{68}$$

using the signed power operator as in the case of (44).

For the optimal filter $\mathbf{K}^{\mathrm{L}}$ so obtained, positive definitiveness of the Hessian matrix

$$\begin{aligned}&\frac{\partial^2}{\partial(K_{ij}^{\mathrm{L}})\partial(K_{iq}^{\mathrm{L}})}\operatorname{trace}\mathbf{B}^{\mathrm{a}}\\ &= \frac{\partial^2}{\partial(K_{ij}^{\mathrm{L}})\partial(K_{iq}^{\mathrm{L}})}B_{ii}^{\mathrm{a}}\\ &= \mu(\mu-1)\left[\sum_{p=1}^{N}\left|G_{ip}^{\mathrm{f}} - \sum_{m=1}^{L}K_{im}^{\mathrm{L}}G_{mp}^{\epsilon}\right|^{\mu-2}(H_{jp}^{\mathbf{G}})(H_{qp}^{\mathbf{G}})C_p^{\mathrm{f}} + \sum_{q=1}^{L}\left|\sum_{m=1}^{L}K_{im}^{\mathrm{L}}G_{mq}^{\epsilon}\right|^{\mu-2}(G_{jp}^{\epsilon})G_{qp}^{\epsilon}C_q^{\epsilon}\right],\end{aligned} \tag{69}$$

is satisfied for $\mu > 1$, following the same approach as discussed in Section 3.2 and in Appendix B.

The optimal estimate $\mathbf{x}_k^{\mathrm{a}}$ and tail-covariance $\mathbf{B}_k^{\mathrm{a}}$ obtained by substituting $\mathbf{K}_k^{\mathrm{L}}$ into (64) and (65) become a set of initial conditions for the next assimilation cycle $k+1$.

Similar to the case of the multivariate estimator which is solution of (44) discussed in Section 3.2, the first condition for the optimal KL filter (68) for a fixed i leads to a set of N self-contained nonlinear equations for N unknowns for $j = 1, \ldots, N$. Minimization of the average scale factor is therefore equivalent to the minimization of each tail-covariance element B_{ii}^{a} for $x_i^{\mathrm{a}} - x_i^{\mathrm{t}}$ with respect to the elements of $\mathbf{K}_k$ with at least one of the indexes equal to i. Such a set of solutions of the nonlinear equations for an arbitrary μ may not be available analytically but can be obtained numerically. For $\mu = 2$, the KL filter is reduced to the same formula as the conventional KG filter, i.e., the forecast error tail-covariance (63) is

$$\mathbf{B}_k^{\mathrm{f}} = \mathbf{M}_{k,k-1}\mathbf{B}_{k-1}^{\mathrm{a}}\mathbf{M}_{k,k-1}^{\mathrm{T}} + \mathbf{B}_{k-1}^{\eta}. \tag{70}$$

The analysis error tail-covariance (65) is

$$\mathbf{B}_k^{\mathrm{a}} = (\mathbf{I} - \mathbf{K}_k^{\mathrm{G}}\mathbf{H}_k)\mathbf{B}_k^{\mathrm{f}}, \tag{71}$$

where the optimal KG gain that satisfies (60) is given by

$$\mathbf{K}_k^{\mathrm{G}} = \mathbf{B}_k^{\mathrm{f}}\mathbf{H}_k^{\mathrm{T}}(\mathbf{H}_k\mathbf{B}_k^{\mathrm{f}}\mathbf{H}_k^{\mathrm{T}} + \mathbf{B}_k^{\epsilon})^{-1}. \tag{72}$$

The analysis state variable $\mathbf{x}_k^{\mathrm{a}}$ is also obtained by substituting (72) into (64). In this case, sequential data assimilation does not require any diagonalization of the tail-covariance matrix.

4.3. Univariate KL filter

4.3.1. Solution

To understand the fundamental properties of the KL filter, we study the univariate problem with $N = 1$ and $L = 1$ in detail for the case where the exponent μ is larger than 1. The case $\mu \leq 1$ has been discussed in Section 2 and leads to a Kalman weight equal to either 0 or 1.

In this case, the tail-covariances $B^{\mathrm{f,a,t}}$ correspond to the scale factors $C^{\mathrm{f,a,t}}$ directly and we have the following data assimilation cycle.

Step 1.

Dynamic forecast:

$$x_k^{\mathrm{f}} = M_{k,k-1}x_{k-1}^{\mathrm{a}}, \tag{73}$$

with tail-covariance matrix

$$B_k^{\mathrm{f}} = |M_{k,k-1}|^{\mu}B_{k-1}^{\mathrm{a}} + B_{k-1}^{\eta}, \tag{74}$$

as derived from (54) using the calculation rules given above in points (1)–(4).

Step 2.

Probabilistic analysis:

$$x_k^{\mathrm{a}} = (1 - K_kH_k)x_k^{\mathrm{f}} + K_ky_k^{\mathrm{o}}, \tag{75}$$

leading to the following scale factor:

$$B_k^{\mathrm{a}} = |1 - K_kH_k|^{\mu}B_k^{\mathrm{f}} + |K_k|^{\mu}B_k^{\epsilon}. \tag{76}$$

Its minimization with respect to K_k leads to

$$x_k^{\mathrm{a}} = \frac{(\lambda_k^H)^{\mu/(\mu-1)}}{1+(\lambda_k^H)^{\mu/(\mu-1)}} x_k^{\mathrm{f}} + \frac{H_k^{-1}}{1+(\lambda_k^H)^{\mu/(\mu-1)}} y_k^{\mathrm{o}}, \tag{77}$$

$$B_k^{\mathrm{a}} = \frac{(\lambda_k^H)^{\mu}}{[1+(\lambda_k^H)^{\mu/(\mu-1)}]^{\mu-1}} B_k^{\mathrm{f}}, \tag{78}$$

after substituting the optimal KL gain

$$K_k^{\mathrm{L}} = \frac{H_k^{-1}}{1+(\lambda_k^H)^{\mu/(\mu-1)}} \tag{79}$$

with the modified relative error ratio

$$\lambda_k^H \equiv \frac{\lambda_k}{H_k} = \frac{(B_k^{\epsilon})^{1/\mu}}{H_k (B_k^{\mathrm{f}})^{1/\mu}}. \tag{80}$$

Notice that expression (78) is nothing but rewriting (14).

4.3.2. Properties of the solution

Because the KL filter is designed to minimize B_k^{a}, we gain insight into its performance by investigating B_k^{a} along with B_k^{f} in each assimilation cycle. While the state variable $x_k^{\mathrm{f,a}}$ has a stochastic dynamics through the observation y^{o}, their scale factors $B_k^{\mathrm{f,a}}$ are completely deterministic. Using (74) and (78), the evolution of $B_k^{\mathrm{f,a}}$ can be expressed as uncoupled one-dimensional maps

$$B_k^{\mathrm{f}} = \frac{|M_{k,k-1}\lambda_{k-1}^H|^{\mu}}{[1+(\lambda_{k-1}^H)^{\mu/(\mu-1)}]^{\mu-1}} B_{k-1}^{\mathrm{f}} + B_{k-1}^{\eta}, \tag{81}$$

$$B_k^{\mathrm{a}} = \frac{(\lambda_k^H)^{\mu}}{[1+(\lambda_k^H)^{\mu/(\mu-1)}]^{\mu-1}} [|M_{k,k-1}|^{\mu} B_{k-1}^{\mathrm{a}} + B_{k-1}^{\eta}], \tag{82}$$

where λ_k^H in B_k^{a} can be given in terms of B_{k-1}^{a} for (82) by substitution of (74) into (80).

Two limiting cases can be analyzed. First, in the limit where the main origin of variability in the factor F multiplying B_{k-1}^{f} in the right-hand side of (81) comes from either $M_{k,k-1}$, B_k^{ϵ} or H_k and not from B_k^{f}, this factor F can be considered to be approximately independent of B_{k-1}^{f}. The expression (81) becomes a multiplicative noisy auto-regressive equation which has been much studied in the literature [5,15,22,23]. The most important result is that B_k^{f} remains finite at all times if the expectation of the logarithm of the factor F multiplying B_{k-1}^{f} in (81) is negative. This condition ensures that B_k^{f} does not grow exponentially at large times. Usually, in this regime, if the factor F exhibits intermittent excursions to values larger than one, it can be shown that the scale factors B_{k-1}^{f} themselves will be distributed according to a power law distribution. A similar result holds for B_k^{a}, whose upper limit is bounded by B_k^{f}.

The second interesting case occurs when the system is stationary, i.e., $M = M_{k,k-1}$, $H = H_k$, and $B^{\eta,\epsilon} = B_k^{\eta,\epsilon}$ for all k. By defining non-dimensional tail-covariances (which are replaced by scale factors in this single variable case) normalized by the dynamical error's scale factor B^{η},

$$b_k^{\mathrm{f,a}} = \frac{B_k^{\mathrm{f,a}}}{B^{\eta}}, \tag{83}$$

the corresponding dynamical maps (81) and (82) are reduced into

$$b_k^{\mathrm{f}} = \frac{|M\lambda^b|^\mu}{\{1+[\lambda^b/(b_{k-1}^{\mathrm{f}})^{1/\mu}]^{\mu/(\mu-1)}\}^{\mu-1}} + 1, \tag{84}$$

$$b_k^{\mathrm{a}} = \frac{(\lambda^b)^\mu}{\{1+[\lambda^b/(|M|^\mu b_{k-1}^{\mathrm{a}}+1)^{1/\mu}]^{\mu/(\mu-1)}\}^{\mu-1}}. \tag{85}$$

For a given μ, the two parameters controlling the evolution of $b_k^{\mathrm{f,a}}$ are the dynamical coefficient M and the ratio of the characteristic error sizes of the dynamics over the observation defined by

$$\lambda^b \equiv \frac{1}{H}\frac{(B^\epsilon)^{1/\mu}}{(B^\eta)^{1/\mu}}. \tag{86}$$

The corresponding KL gain is

$$K_k^{\mathrm{L}} = \frac{H^{-1}}{1+[\lambda^b/(b_{k-1}^{\mathrm{f}})^{1/\mu}]^{\mu/(\mu-1)}}, \tag{87}$$

which can be expressed in terms of b_{k-1}^{a} as well. The KL filter parameter (86) and gain (87) are the counterpart of (12) and (13), and the resulting KL analysis (85) takes the form similar to (14) of the Lévy estimator in Section 2.2.

For any values of $b_{k-1}^{\mathrm{f,a}}$, the scale factors $b_k^{\mathrm{f,a}}$ at the next time step are bounded:

$$1 < b_k^{\mathrm{f}} < |M\lambda^b|^\mu + 1, \tag{88}$$

$$\frac{(\lambda^b)^\mu}{[1+(\lambda^b)^{\mu/(\mu-1)}]^{\mu-1}} < b_k^{\mathrm{a}} < (\lambda^b)^\mu, \tag{89}$$

indicating that the univariate sequential estimation system cannot diverge as long as the exponent μ is finite.

The evolution of the scale factors is demonstrated in Fig. 4a. The maps are represented by the graphs of b_k^{f} and b_k^{a} as functions of b_{k-1}^{a} for four values of the exponent $\mu = 1.2, 1.5, 2$ and 3 with the set of parameters $(\lambda^b, M) = (1, 0.9)$, corresponding to a contracting map. Starting from b_{k-1}^{a} on the diagonal line, the dynamic forecast takes b_{k-1}^{a} to b_k^{f} on the corresponding μ-curve (upper group), which is followed by the probabilistic analysis down to b_k^{a} on the corresponding μ-curve (lower group) to complete one data assimilation cycle, $b_{k-1}^{\mathrm{a}} \to b_k^{\mathrm{f}} \to b_k^{\mathrm{a}}$. A new analysis b_k^{a} is moved horizontally onto the diagonal line to become an initial scale factor value for the next cycle.

On each b_k^{a} curve for a fixed μ, the symbols (circle, square, diamond and triangle for $\mu = 1.2, 1.5, 2$ and 3, respectively) at the intersection with the diagonal line, i.e. $\bar{b}^{\mathrm{a}} = b_k^{\mathrm{a}} = b_{k-1}^{\mathrm{a}}$, are the stable solutions of the KL filter. The maps (84) and (85) in fact have each a stable fixed-point solution, $\bar{b}^{\mathrm{f}}$ and $\bar{b}^{\mathrm{a}}$ that attracts any initial condition for any exponent μ, given a set of parameters (λ^b, M). For $\mu = 2$ retrieving the case of the KG filter, this stable fixed-point can be obtained analytically [9].

The error probability distributions for the stable fixed-points corresponding to the Student's distribution (24) are shown in Fig. 4b–e based on the scale factor $\bar{b}^{\mathrm{f,a}}$. We use $(x^{\mathrm{f}}, y^{\mathrm{o}}) = (0, 1)$ and $H = 1$, so that $x^{\mathrm{a}} = K^{\mathrm{L}}$. For small μ, the KL filter favors more strongly the better sample characterized by the smaller scale factor between the two ($\bar{b}^{\mathrm{f}}$ for x^{f} and $(\lambda^b)^\mu$ for y^{o}). This effect is stronger when μ decreases to 1, as discussed in Section 2. The value of the fixed-point $\bar{b}^{\mathrm{a}}$ is larger for smaller μ, i.e., a system with a probability distribution with heavier tail has greater uncertainty, not only because of its slow decay measured by the exponent μ, but also due to its overall amplitude quantified by the scale factor. Furthermore, the slope of the curve b_k^{a} as a function of b_{k-1}^{a} shown in Fig. 4 is closer to the horizontal for smaller μ, indicating that the convergence to the stable fixed-point is faster for the heavy tail probability distribution. This is because the KL filter with smaller μ tends to favor either forecast or observation

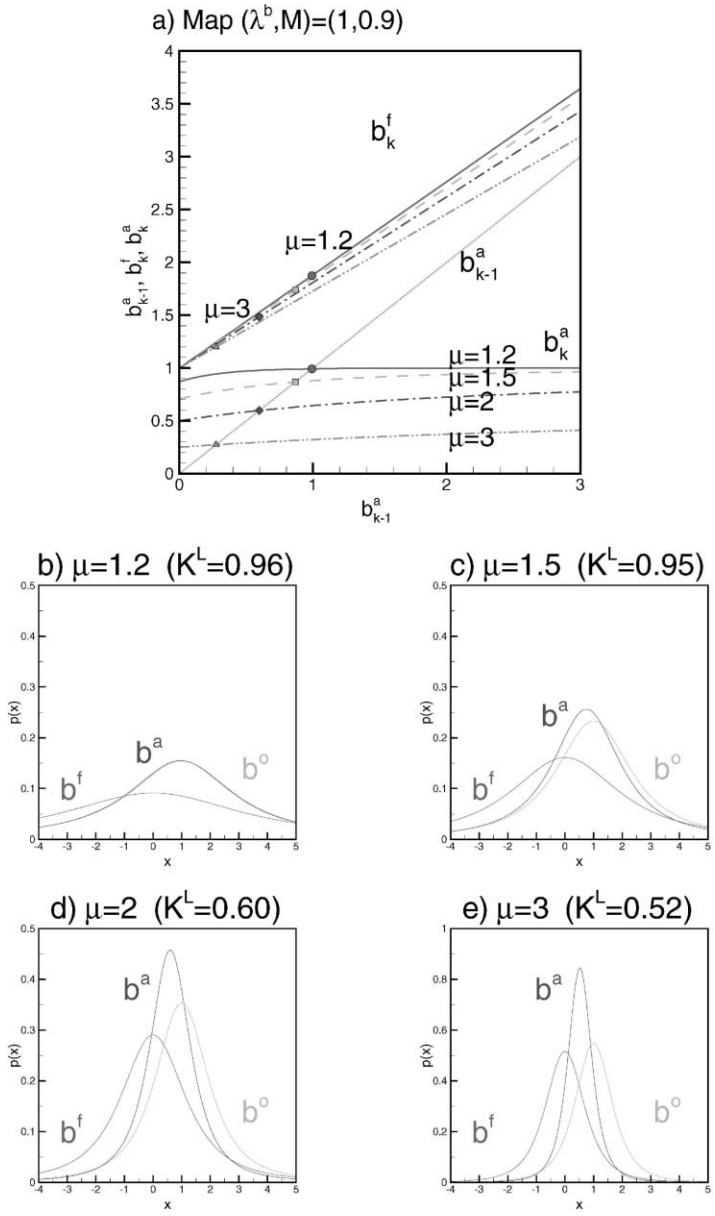

Fig. 4. (a) Graphs of b_k^{f} and b_k^{a} as a function of b_{k-1}^{a} for $(\lambda^b, M) = (1, 0.9)$. For both b_k^{f} and b_k^{a}, the lines represent the corresponding maps and the symbols are the stable fixed-points; solid, dash, dash-dot, and dash-dot-dot lines, as well as circle, square, diamond, and triangle symbols correspond to $\mu = 1.2$, 1.5, 2 and 3, respectively. (b) Error probability distribution for the stable fixed-point corresponding to $\mu = 1.2$ and parameters $(x^{\mathrm{f}}, y^{\mathrm{o}}) = (0, 1)$ and $H = 1$, so that $x^{\mathrm{a}} = K^{\mathrm{L}}$. (c)–(e) Same as (b) but corresponding to $\mu = 1.5$, 2, and 3, respectively.

strongly, depending on their relative noise amplitude quantified by the scale factor of their noises as discussed in Section 2.3.

Since the stationary KL assimilation system quickly approaches a unique steady state for a given set of parameters (λ^b, M) and exponent μ, the stable fixed-point defined by $\bar{b}^{\mathrm{f}}$ and $\bar{b}^{\mathrm{a}}$ along with the corresponding optimal KL gain

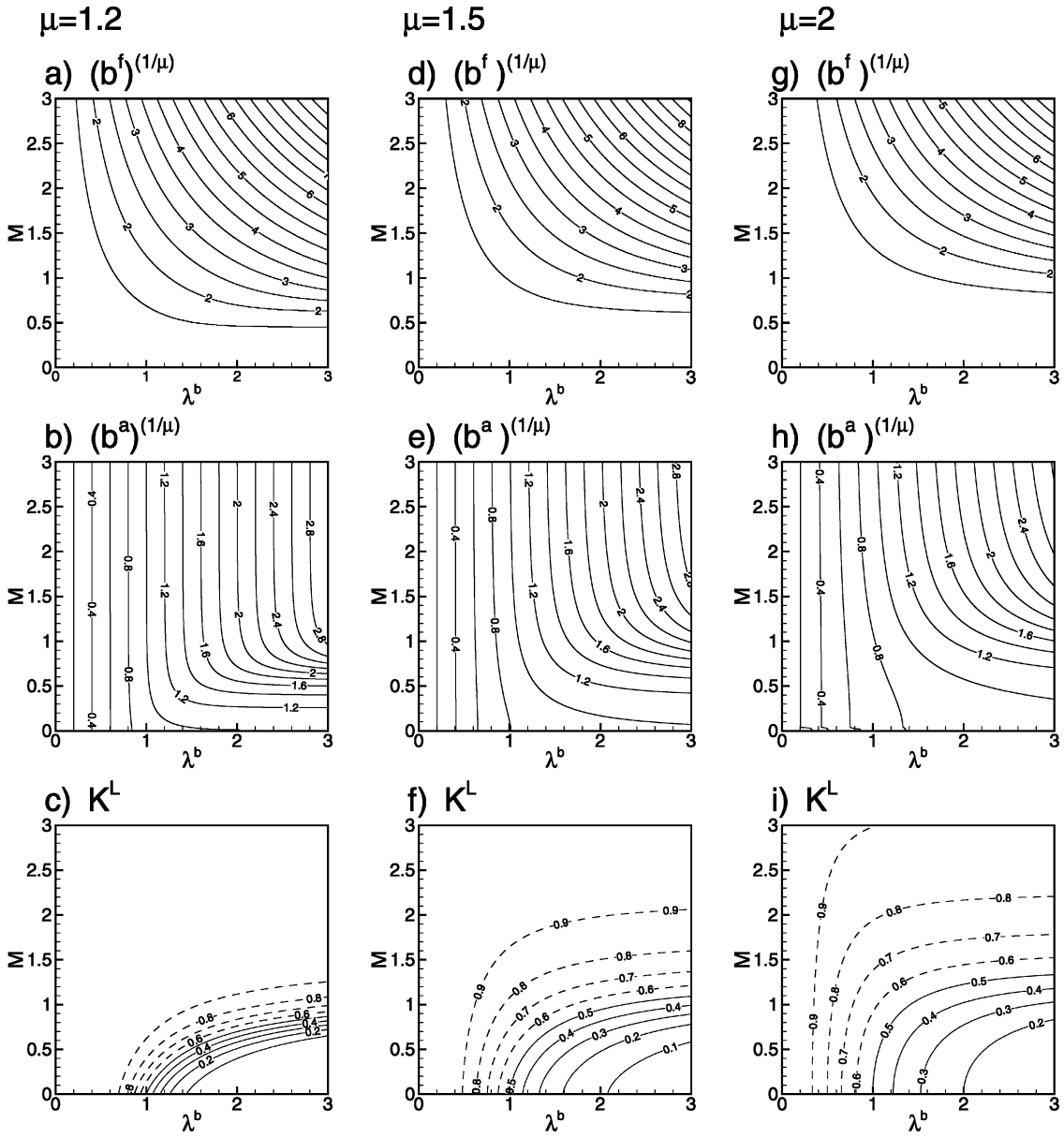

Fig. 5. Stable fixed-point of the KL filter as a function of λ^b and M: (a) $(\bar{b}^{\mathrm{f}})^{1/\mu}$; (b) $(\bar{b}^{\mathrm{a}})^{1/\mu}$; (c) $\bar{K}^{\mathrm{L}}$ for $\mu = 1.2$; (d)–(f) same as (a)–(c) but for $\mu = 1.5$; (g)–(i) same as (a)–(c) but for $\mu = 2$.

$\bar{K}^{\mathrm{L}}$ suffice to provide a complete description of the stationary assimilation process. Fig. 5 shows the stable solution obtained from (84), (85) and (87), which are plotted in the parameter space (λ^b, M) for $\mu = 1.2$, 1.5 and 2. All tail-covariances (scale factors) are plotted in terms of $(\bar{b})^{1/\mu}$ so as to preserve the characteristic scale as in (6) of the error, independently of the value of μ.

For convergent dynamics $M < 1$ with sufficiently large relative observational error $\lambda^b > 1$, $\lambda^H \gg 1$ and hence $\bar{K}^{\mathrm{L}} \approx 0$ favors the forecast (Fig. 5c, f and i) and hence results in $\bar{b}^{\mathrm{a}} \sim \bar{b}^{\mathrm{f}}$. For divergent dynamics with a larger

value of $M > 1$, however, it yields a large value for the forecast's scale factor $\bar{b}^{\mathrm{f}}$ (Fig. 5a, d and g) with respect to relative observational error λ^b. Here, the KL filter correctly derives that the errors are amplified by the unstable evolution of the system. Since $\bar{b}^{\mathrm{f}}$ is large and hence $\lambda^H \approx 0$, $\bar{K}^{\mathrm{L}} \approx 1$ favors the observation over the forecast and hence results in $\bar{b}^{\mathrm{a}} \sim (\lambda^b)^\mu$ (Fig. 5b, e and h). This effect derived by the KL filter is more significant for heavy tail probability distributions with a smaller μ, and it is manifested in the much steeper gradient of $\bar{K}^{\mathrm{L}}$ for $\mu = 1.2$ (Fig. 5c) in contrast to the widely spread contour maps observed for $\mu = 2$ (Fig. 5i).

Note that for $M = 0$, we retrieve the univariate filter studied in Section 2. In particular, for $\lambda^b = 1$, $K^{\mathrm{L}} = \frac{1}{2}$, corresponding to equal weights of the assimilation on observation and forecast for any exponent μ. The curvature to the right taken by the contour maps of the Kalman gain K as a function of M can easily be rationalized: a larger M implies a larger forecast error, hence a smaller effective λ^b. One thus needs a large observation over dynamical error ratio to get the same effective effect, hence the downward convexity of the contour maps.

4.3.3. Relative performance of the KL and KG filters

We now examine the case where the KG filter K^{G} (corresponding to putting $\mu = 2$ in the solutions) is applied to the system whose true noise distribution is the heavy tail power law ($\mu < 2$). This may happen in a practical implementation of Kalman filtering when we do not know the nature of the noises very well and a finite variance assumption is made. This is also probably the only choice left to the operator in absence of our solution presented in this paper. This exercise is thus aimed at quantifying what we have gained concretely by recognizing the non-Gaussian nature of the noise and by providing the corresponding solution. We stress that what matters according to the KG framework is whether the noise has a finite variance or not. In other words, all non-Gaussian noises with finite variance are treated in the same fashion within the KG approach by analyzing the variance only. In contrast, the KL solution distinguishes noises even if they have a finite variance by analyzing the structure of their "fat-tail" characterized by the exponent μ. For instance, the KL gain is different for noise distributions with power law tails with $\mu = 3$ and 4, while in contrast the KG solution is the same for both if they have the same variance.

We formulate this scenario in a general form when an incorrect model exponent $\tilde{\mu}$ is used to assimilate the data from the system with true exponent μ. In case of the KG filter application, this means $\tilde{\mu} = 2$. The model for data assimilation that we obtain is equivalent to (73)–(80) by replacing

$$(x_k^{\mathrm{f,a}}, B_k^{\mathrm{f,a}}, K_k) \to (\tilde{x}_k^{\mathrm{f,a}}, \tilde{B}_k^{\mathrm{f,a}}, \tilde{K}_k), \tag{90}$$

$$[\mu, B_k^{\eta,\epsilon}, H\lambda_k^H] \to [\tilde{\mu}, (\tilde{B}_k^{\eta,\epsilon})^{\mu/\tilde{\mu}}, (H\tilde{\lambda}_k^H)^{\mu/\tilde{\mu}}], \tag{91}$$

where $\{\tilde{\cdot}\}$ represents the model filtering. The exponent factor $\{\cdot\}^{\mu/\tilde{\mu}}$ arises so as to preserve the characteristic scale of the noises.

Use of the model gain $\tilde{K}_k \neq K_k^{\mathrm{L}}$ due to an incorrect model exponent $\tilde{\mu}$ yields a non-optimal filtering by definition, in a sense that the analysis scale factor B_k^{a} is not a minimum. In addition, such a model filtering estimates both forecast and analysis tail-covariances $\hat{B}_k^{\mathrm{f,a}}$ incorrectly as $\tilde{B}_k^{\mathrm{f,a}}$, because the real evolution of the tail-covariances using the non-optimal model gain $\tilde{K}_k$ should follow the non-optimal KL filtering scheme which itself uses the true exponent μ

$$\hat{B}_k^{\mathrm{f}} = |M_{k,k-1}|^\mu \hat{B}_{k-1}^{\mathrm{a}} + B_{k-1}^{\eta}, \tag{92}$$

$$\hat{B}_k^{\mathrm{a}} = |1 - \tilde{K}_k H_k|^\mu \hat{B}_k^{\mathrm{f}} + (\tilde{K}_k)^\mu B_k^{\epsilon}. \tag{93}$$

Accordingly, there are three filtering representations, i.e.:

(1) true and optimal KL filtering $B_k^{\mathrm{f,a}}$ using $(\mu, \bar{K}^{\mathrm{L}})$;
(2) true but non-optimal filtering $\hat{B}_k^{\mathrm{f,a}}$ using $(\mu, \tilde{K}^{\mathrm{G}})$;
(3) model and incorrect filtering $\tilde{B}_k^{\mathrm{f,a}}$ using $(\tilde{\mu}, \tilde{K}^{\mathrm{G}})$.

Table 1
Stable fixed-points of (1) optimal KL, (2) non-optimal KG and (3) incorrect KG, using the exponents $(\mu, \tilde{\mu}) = (1.2, 2)$ and system parameter set $(\lambda^b, M) = (1, 0.9)$

	Optimal KL $(\bar{b})(\mu, \bar{K}^{\mathrm{L}})$	Non-optimal $(\hat{b})(\mu, \tilde{K}^{\mathrm{G}})$	Model $(\bar{\tilde{b}})(\tilde{\mu}, \tilde{K}^{\mathrm{G}})$
b^{f}	1.87	2.09	1.48
b^{a}	0.99	1.24	0.59
K	0.96	0.60	0.60

By definition, the optimal filtering (1) is always superior to the non-optimal filtering (2) $(B_k^{\mathrm{a}})^{1/\mu} \leq (\hat{B}_k^{\mathrm{a}})^{1/\mu}$. It is possible, however, that the incorrect model filtering (3) returns a value for the scale factor which is numerically smaller (see Table 1). Since the model exponent $\tilde{\mu}$ is different from the true exponent μ, the scale factors cannot be compared directly to infer the quality of the assimilation process.

When the system is time-independent, the normalized one-dimensional maps of the tail-covariances using the non-optimal KL filter with model gain $\tilde{K}$ are

$$\hat{b}_k^{\mathrm{f}} = |M(1 - \tilde{K}H)|^{\mu} \hat{b}_{k-1}^{\mathrm{f}} + |M\tilde{K}H\lambda^b|^{\mu} + 1, \tag{94}$$

$$\hat{b}_k^{\mathrm{a}} = |M(1 - \tilde{K}H)|^{\mu} \hat{b}_{k-1}^{\mathrm{a}} + |1 - \tilde{K}H|^{\mu} + (\tilde{K}H\lambda^b)^{\mu}. \tag{95}$$

This non-optimal filtering also has a unique stable fixed-point

$$\bar{\hat{b}}^{\mathrm{f}} = \frac{|M\tilde{K}H\lambda^b|^{\mu} + 1}{1 - |M(1 - \tilde{K}H)|^{\mu}}, \tag{96}$$

$$\bar{\hat{b}}^{\mathrm{a}} = \frac{|1 - \tilde{K}H|^{\mu} + (\tilde{K}H\lambda^b)^{\mu}}{1 - |M(1 - \tilde{K}H)|^{\mu}}, \tag{97}$$

provided the condition for stability is satisfied

$$0 < |M(1 - \tilde{K}H)| < 1. \tag{98}$$

To see this effect of non-optimal filtering for the KG filter application, we apply the model gain $\tilde{K} = \tilde{K}^{\mathrm{G}}$ with $\mu = 2$ (Fig. 5i) to a time independent system (84)–(87) subjected to the Lévy noise with true exponent $\mu = 1.2$. The stable assimilation cycle for this optimal KL filtering have been presented in Fig. 5a–c. The unique stable fixed-points of the non-optimal filter given by (96) and (97) are shown in Fig. 6a and b, in terms of the characteristic scale $(\bar{\hat{b}}^{\mathrm{f,a}})^{1/\mu}$. Because the KG filtering is no longer optimal, $\bar{\hat{b}}^{\mathrm{f,a}}$ are now larger than the corresponding optimal scale factors $\bar{b}^{\mathrm{f,a}}$ of the KL fixed-point (Fig. 5a and b).

To quantify the difference between the KL and KG solutions, we construct the differences of the normalized stable fixed-point found in the three assimilation representations (1)–(3). In Fig. 6c–f, we present the comparison for the following two cases:

1. difference between the non-optimal filtering ((2) as in Fig. 6a and b) and optimal KL filtering ((1) as in Fig. 5a and b);
2. difference between non-optimal filtering (2) and incorrect model filtering (3).

All results are shown in terms of the characteristic error scales, $b^{1/\mu}$ or $b^{1/\tilde{\mu}}$, so that the comparison can be made independent of the exponents μ and $\tilde{\mu}$ in the filters.

The first comparison between (2) and (1) corresponds to the difference between the optimal and non-optimal filtering. In Fig. 6c and d, we observe a bimodal structure in the difference $(\hat{b}^{\mathrm{f,a}})^{1/\mu} - (\bar{b}^{\mathrm{f,a}})^{1/\mu}$, caused by the

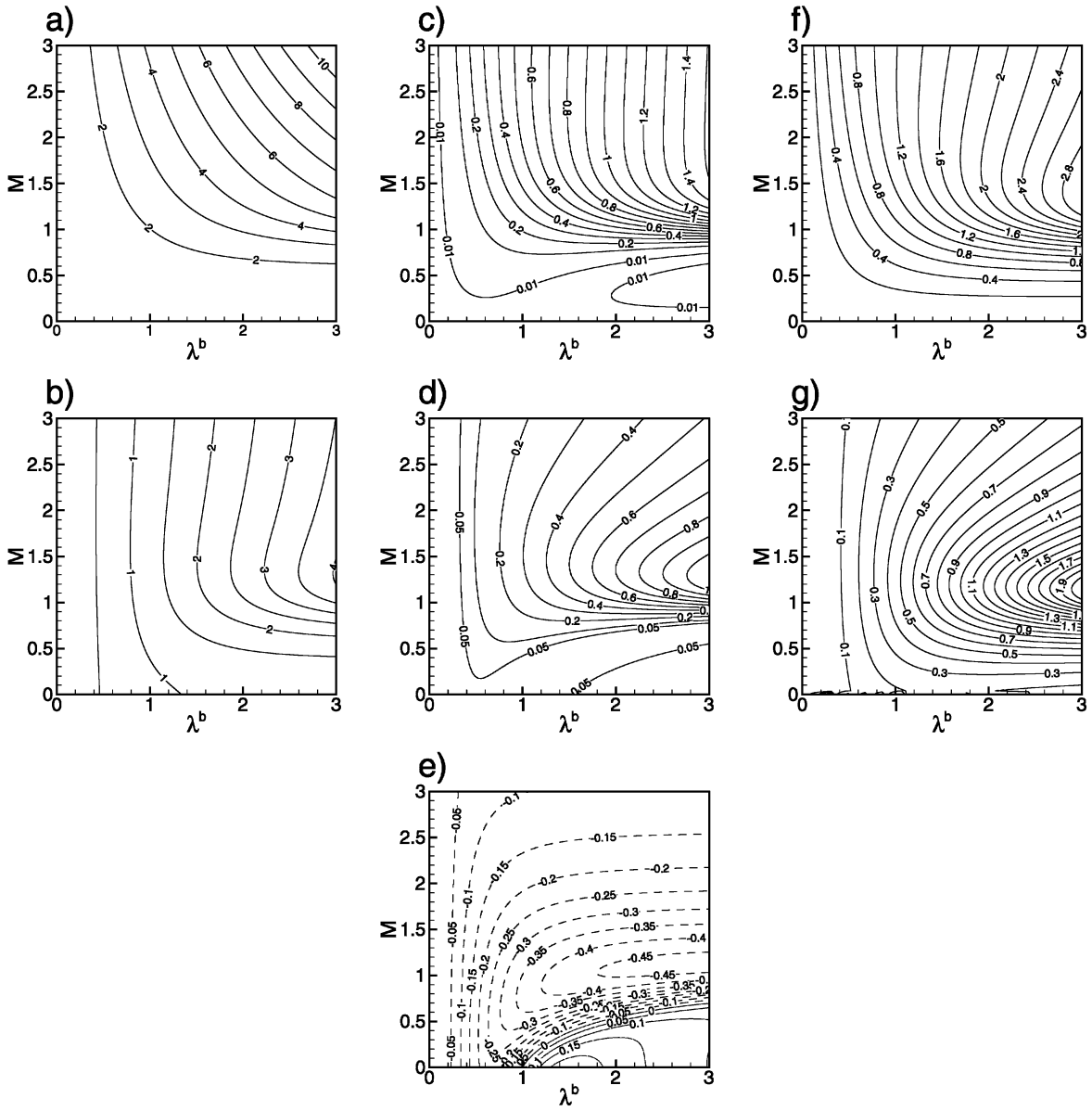

Fig. 6. Stable fixed-point of the KG filter applied to a heavy tail system with $\mu = 1.2$: (a) $(\hat{b}^{\mathrm{f}})^{1/\mu}$; (b) $(\hat{b}^{\mathrm{a}})^{1/\mu}$; (c) difference $(\bar{\hat{b}}^{\mathrm{f}})^{1/\mu} - (\bar{b}^{\mathrm{f}})^{1/\mu}$ between the non-optimal filtering solution $(\bar{\hat{b}}^{\mathrm{f}})^{1/\mu}$ shown in (a) and the optimal KL filtering solution $(\bar{b}^{\mathrm{f}})^{1/\mu}$ shown in Fig. 5a; (d) difference $(\hat{b}^{\mathrm{a}})^{1/\mu} - (\bar{b}^{\mathrm{a}})^{1/\mu}$ between the non-optimal filtering solution $(\hat{b}^{\mathrm{a}})^{1/\mu}$ shown in (b) and the optimal KL filtering solution $(\bar{b}^{\mathrm{a}})^{1/\mu}$ shown in Fig. 5b; (e) difference $K^{\mathrm{G}} - K^{\mathrm{L}}$ between the non-optimal filtering solution and the model filtering solution; (f) difference $(\hat{b}^{\mathrm{f}})^{1/\mu} - (\tilde{b}^{\mathrm{f}})^{1/\tilde{\mu}}$ between the non-optimal filtering solution $(\hat{b}^{\mathrm{f}})^{1/\mu}$ shown in (b) and the model filtering solution $(\tilde{b}^{\mathrm{f}})^{1/\tilde{\mu}}$; (g) difference $(\hat{b}^{\mathrm{a}})^{1/\mu} - (\tilde{b}^{\mathrm{a}})^{1/\tilde{\mu}}$ between the non-optimal filtering solution $(\hat{b}^{\mathrm{a}})^{1/\mu}$ shown in (b) and the mode filtering solution $(\tilde{b}^{\mathrm{a}})^{1/\tilde{\mu}}$.

maximum–minimum structure in the gain $\tilde{K}^{\mathrm{G}} - \bar{K}^{\mathrm{L}}$ (Fig. 6e), whose origin is the following. For $M < 1$ for which $M^2 < M^\mu$ for $\mu < 2$, $\tilde{K}^{\mathrm{G}} > \bar{K}^{\mathrm{L}}$. The non-optimal gain $\tilde{K}^{\mathrm{G}}$ obtained from the model KG solution thus overestimates the uncertainty of the forecast. On the other hand, for $M > 1$ for which $M^2 > M^\mu$ for $\mu < 2$, $\tilde{K}^{\mathrm{G}} < \bar{K}^{\mathrm{L}}$. The non-optimal gain $\tilde{K}^{\mathrm{G}}$ now underestimates the reliability of the observation. Both ways, it increases the non-optimal solution $\hat{b}^{\mathrm{a}}$ in comparison to the optimal solution $\bar{b}^{\mathrm{a}}$.

The second comparison between (2) and (3) relates to the actual mistake in the model solution $\tilde{b}^{\mathrm{f,a}}$ made by using the model gain $\tilde{K}^{\mathrm{G}}$ to the system which reaches a different, non-optimal stable fixed-point $\hat{b}^{\mathrm{f,a}}$. As shown in 6f and g $(\hat{b}^{\mathrm{f,a}})^{1/\mu} - (\tilde{b}^{\mathrm{f,a}})^{1/\tilde{\mu}}$ are positive and therefore the model filtering incorrectly underestimates the error.

Fig. 7 is the KG filtering application when the true exponent $\mu = 1.5$ is not as heavy as the previous case $\mu = 1.2$. Although the KL solution is better than the KG one as expected, the difference is smaller: the improvement is of the order of 5–10% at most. The fact that the improvement has a smaller amplitude is clear: a larger exponent implies a thinner tail and thus a behavior closer to the Gaussian case. Recall that at μ goes to 2, the Gaussian case and solution are recovered.

4.3.4. Numerical experiment

To check these results derived from the analysis of the deterministic behavior of the tail-covariances (scale factors) $b_k^{\mathrm{f,a}}$, we present a numerical experiment of the stochastic dynamics, observation construction and assimilation processes. We use the parameter set $(\lambda^b, M) = (1, 0.9)$ with noises distributed according to a Lévy law with exponent $\mu = 1.2$. The one-dimensional map and probability distribution of the stable fixed-point for this parameter set are shown in Fig. 4a and b. To generate the Lévy noises, we follow the standard algorithm described initially in [4] and use the software available in [17]. The stochastic variables x_k^{t} and y_k^{o} are generated over 10 000 time steps.

Typical results of the KL filtering $(\mu, \bar{K}_k^{\mathrm{L}})$ are shown in Fig. 8. The true evolution x_k^{t} is shown over the 10 000 time steps in Fig. 8a. Note the occurrence of a few very large fluctuations that dwarf most of the remaining dynamics. To get a closer view, we enlarge Fig. 8a in the narrow time interval [1000, 1050] shown in Fig. 8b and c. Fig. 8b gives the dynamical evolution of the true, forecasted, observation and analysis variables, when using the optimal KL filtering, while Fig. 8c corresponds to the use of the non-optimal filtering. Note that the two filtering's use the same gain $\tilde{K}_k^{\mathrm{G}}$ and therefore result in the same $\tilde{x}_k^{\mathrm{f,a}}$. One can observe on Fig. 8b that the optimal KL filtering x_k^{a} follows rather closely the observation y_k^{o}. This results from the high value of $\bar{K}_k^{\mathrm{L}}$. In contrast, the non-optimal filtering puts $\tilde{x}_k^{\mathrm{a}}$ midway between y_k^{o} and $\tilde{x}_k^{\mathrm{f}}$. The tail-covariances (scale factor) quickly approaches the stable fixed-point after a few iterations as given in Table 1, along with the stable fixed-points of the non-optimal filtering $\hat{b}^{\mathrm{f,a}}$ with $(\mu, \tilde{K}_k^{\mathrm{G}})$ and model filtering $\tilde{b}^{\mathrm{f,a}}$ with $(\tilde{\mu}, \tilde{K}_k^{\mathrm{G}})$.

Because the KL filter is designed for the global control of the uncertainty by minimizing the tail-covariance, we propose to compare the tails of the distribution of errors $(x^{\mathrm{a}} - x^{\mathrm{t}})$ resulting from the two methods (KL and KG) to assess their relative performance. Note that since the covariance does not exist for $\mu < 2$, it cannot be used to evaluate the performance of the heavy tail KL filtering.

Fig. 9a shows the (complementary) cumulative distribution of $(x^{\mathrm{a}} - x^{\mathrm{t}})$ and $(x^{\mathrm{f}} - x^{\mathrm{t}})$, as well as that of $(y^{\mathrm{o}}/H) - x^{\mathrm{t}}$ for reference. For this parameter set $(\lambda^b, M) = (1, 0.9)$, the two optimal KL and model KG gains at their stable fixed-points differ by 37.5% (Table 1). This shows that the model KG filter underestimates the reliability of observation and overestimates the value of the forecast.

Although the difference in the cumulative distributions is rather subtle to determine from visual inspection of Fig. 9a, the cumulative distribution of the error between the analysis and the true trajectory obtained from the optimal KL filter is consistently below that obtained by using the non-optimal filter, apart from expected fluctuations. In probability terms, the optimal KL error distribution exhibit the property of being "stochastically dominant" over the model KG error distribution. This shows that the optimal KL filter is indeed superior to the model KG filter in the presence of heavy tails.

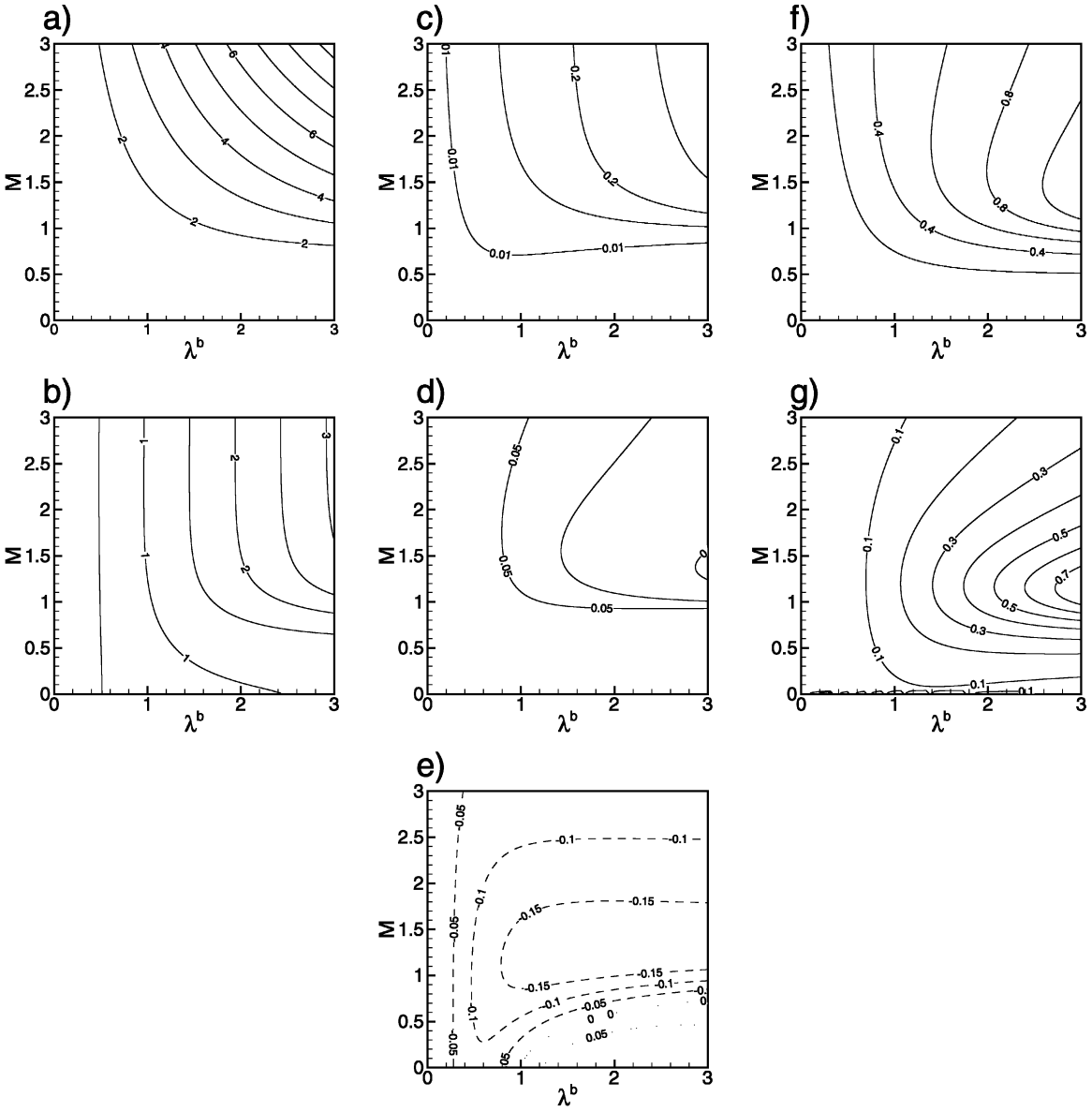

Fig. 7. Same as Fig. 6 for dynamical and observational noises given by Lévy distributions with exponent $\mu = 1.5$.

In fact, our theory predict the difference of 37.5% (Table 1) based on the stable fixed-point presented in Section 4.3.2. It is confirmed by the synthetic simulation of the Lévy random variables with $\mu = 1.2$ as presented in Fig. 9b based on the corresponding scale factors of the stable fixed-point $\bar{b}^{\mathrm{f,a}}$ and $\hat{b}^{\mathrm{f,a}}$ (Table 1).

The superficial visual effect in Fig. 9a can be explained as follows. In the tails, Lévy laws with exponent μ are power law given by $C/(x^{\mathrm{a}} - x^{\mathrm{t}})^{\mu}$, where C is the scale factor of the errors. If C is higher by 37.5% for the non-optimal filtering compared to the optimal KL filtering (Table 1), this represents a significant error reduction.

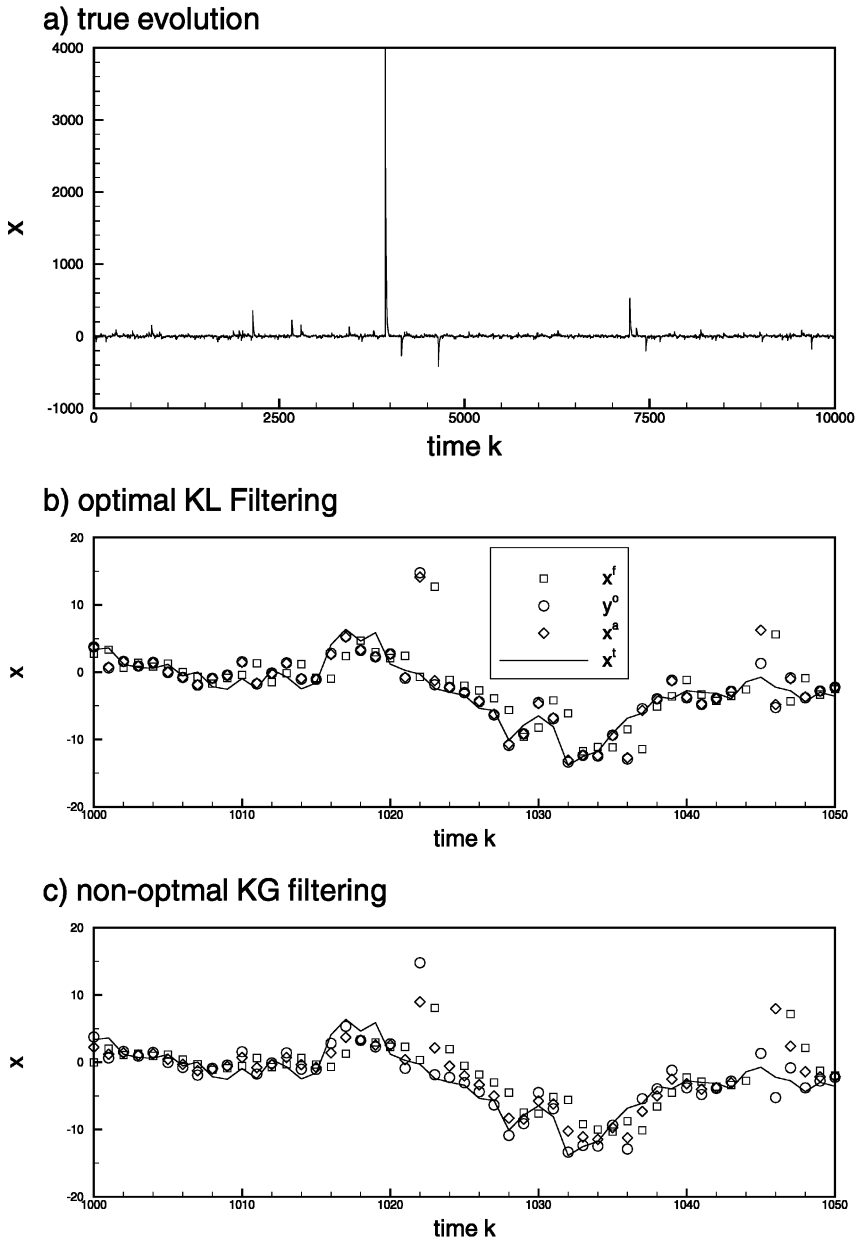

Fig. 8. Numerical experiment for the parameter set $(\lambda^b, M) = (1, 0.9)$ with $\mu = 1.2$: (a) evolution of true state variable x_k^{t} for $k = 1, \ldots, 10\,000$. Note the occurrence of a few large peaks corresponding to rare but extreme noise fluctuations distributed with the Lévy distribution. Panels (b) and (c) show a magnification of panel (a) in the time interval [1000, 1050] and compare this true dynamics (continuous line) with the forecasts x^{f} (squares), the observations y^{o} (circles) and the assimilation analysis x^{a} (diamonds). Panel (b) corresponds to the use of the optimal KL filtering while panel (c) corresponds to the model filtering with $\tilde{\mu} = 2$, i.e. standard KG filter. It appears clear by visual inspection that the optimal KL analysis x^{a} is much closer more often than not to the true dynamics.

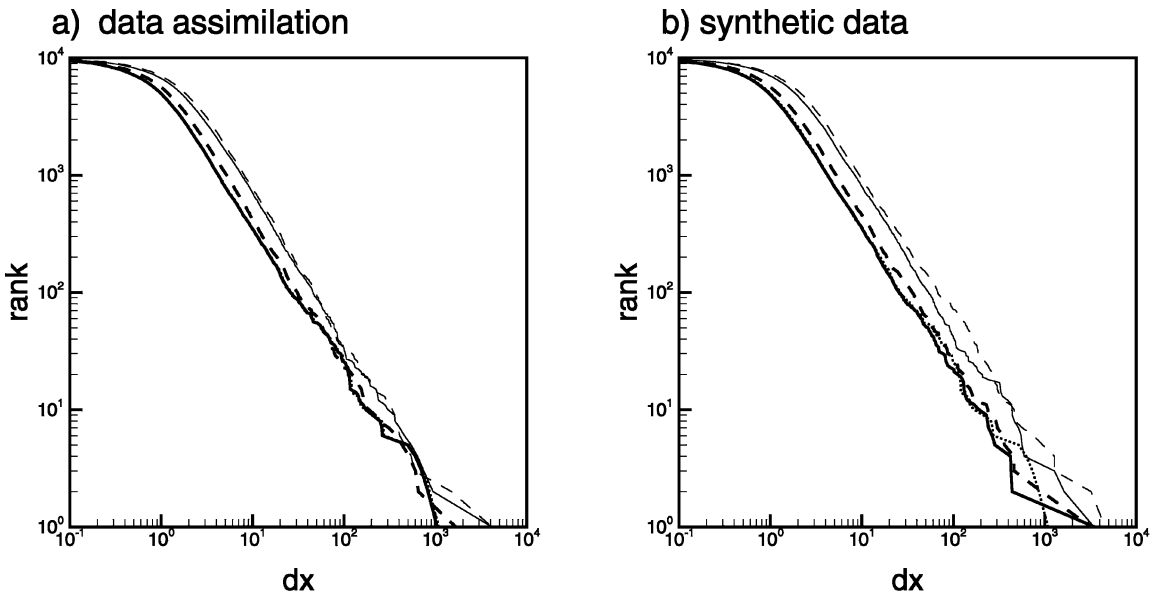

Fig. 9. (a) (Complementary) cumulative distribution (i.e. number of time steps where the error is larger than a value read on the abscissa) of the error between the analysis and the true trajectory for $(\lambda^b, M) = (1, 0.9)$ and $\mu = 1.2$, obtained by using the optimal KL filter K^{L} ($x^{\mathrm{a}} - x^{\mathrm{t}}$; in thick solid line) and model KG filter $\tilde{K}_k$($\tilde{x}^{\mathrm{a}} - x^{\mathrm{t}}$; in thick dashed line). We observe clearly that the distribution of $x^{\mathrm{a}} - x^{\mathrm{t}}$ is below that of $\tilde{x}^{\mathrm{a}} - x^{\mathrm{t}}$, i.e. the errors are globally reduced *in distribution* by application of the KL method compared to the standard KG method. We also show the cumulative distributions of the difference between forecast and true trajectory for the optimal KL filter $K^{\mathrm{L}}(x^{\mathrm{f}} - x^{\mathrm{t}}$; in solid line) and model KG filter $\tilde{K}_k(\tilde{x}^{\mathrm{f}} - x^{\mathrm{t}}$; in dashed line), as well as the cumulative distributions of the observations $((y^{\mathrm{o}}/H) - x^{\mathrm{t}}$; in dotted line, which is almost identical to $x^{\mathrm{a}} - x^{\mathrm{t}}$ and thus hardly visible due to the thickness of the lines). (b) (Complementary) cumulative distribution of the synthetic simulation of random Lévy variables based on the scale-factors of the stable fixed-points $\bar{b}^{\mathrm{f,a}}$ and $\tilde{b}^{\mathrm{f,a}}$ (Table 1) predicted by the theory. The cumulative distributions of the same observations $((y^{\mathrm{o}}/H) - x^{\mathrm{t}}$; in dotted line) as in (a) is also shown for reference. This alternative method for constructive the distributions of errors shows the full consistency of the approach.

However, this will not be strikingly visible in the log–log representation of Fig. 9, since $\ln 1.375 \approx 0.32$ leads to a translation of the two cumulative distributions by only 0.32, hence the small but still visible effect.

Another more compact way of quantifying the relative performance of the two solutions is to calculate a typical error amplitude, which generalizes the covariance. Since we have considered the situation where $\mu > 1$, the average of the *absolute value* $\langle |x^{\mathrm{a}} - x^{\mathrm{t}}| \rangle$ of the errors corresponds to a moment of order 1, which is defined mathematically and is numerically well behaved. Our direct numerical simulations show a decrease of the typical error amplitude ($\langle |x^{\mathrm{a}} - x^{\mathrm{t}}| \rangle$) by approximately 20% when going from the non-optimal solution (≈ 3.3) to the optimal KL filtering (≈ 2.8).

5. Conclusion

We have presented the solution of the Kalman filter problem for dynamical and forecast noises distributed according to power laws and Lévy laws. The main theoretical concept that we have used is to optimize the Kalman filter to chisel the tail of the distribution of residual errors so as to minimize it globally. In order to implement this program, we have introduced the concept of a "tail-covariance" that generalizes the usual notion of the covariance. The full solution, called the KL filter, is obtained by the solution of a general nonlinear equation. We have investigated in detail the quality of this solution in the univariate case and have shown by direct numerical experiments that the improvement is significant, all the more so, the heavier the tail, i.e. the smaller the power law exponent μ.

Acknowledgements

We acknowledge stimulating discussions with Dan Cayan and Larry Riddle at Scripps and Michael Ghil and Andrew Robertson at UCLA and thank R. Mantegna for exchanges on how to generate noise with Lévy distributions. We thank the two referees for spotting analytical errors in the submitted version and for their insightful remarks that helped improve the presentation. All errors remain ours. This work was partially supported by NASA NA65-92 94 ONR N00014-99-1-0020 (KI). D.S. gratefully acknowledges support from the James S. McDonnell Foundation 21st Century Scientist award/studying complex systems.

Appendix A. Stable Lévy laws

The stable laws have been studied and classified by Paul Lévy, who discovered that in addition to the Gaussian law, there is a large number of other p.d.f.'s sharing the stability condition

$$P_N(x')\,\mathrm{d}x' = P_1(x)\,\mathrm{d}x, \quad \text{where } x' = a_N x + b_N \tag{A.1}$$

for some constants a_N and b_N, where x' is the sum of N independent variables of the type x distributed according to the p.d.f. $P_1(x)$. One of their most interesting properties is their asymptotic power law behavior.

A symmetric Lévy law centered on zero is completely characterized by two parameters which can be extracted solely from its asymptotic dependence

$$P(x) \sim \frac{C}{|x|^{1+\mu}} \quad \text{for } x \to \pm\infty. \tag{A.2}$$

C is a positive constant called the tail or scale parameter and the exponent μ is between 0 and 2 ($0 < \mu < 2$). Clearly, $\mu > 0$ for the p.d.f. to be normalizable. As for the other condition $\mu < 2$, a p.d.f. with a power law tail with $\mu > 2$ has a finite variance and thus converges (slowly) in probability to the Gaussian law. It is therefore not stable. Only its shrinking tail for $\mu > 2$ remains of the power law form. In contrast, the whole Lévy p.d.f. remains stable for $\mu < 2$.

All symmetric Lévy law with the same exponent μ can be obtained from the Lévy law $L_\mu(x)$ with the same exponent μ, centered on zero and with unit scale parameter $C = 1$, under translation and rescaling transformations

$$P(x)\,\mathrm{d}x = L_\mu(x')\,\mathrm{d}x', \quad \text{where } x' = C^{1/\mu}x + m, \tag{A.3}$$

m being the center parameter.

Lévy laws can be asymmetric and the parameter quantifying this asymmetry is

$$\beta = \frac{C_+ - C_-}{C_+ + C_-}, \tag{A.4}$$

where $C_\pm$ are the scale parameters for the asymptotic behavior of the Lévy law for $x \to \pm\infty$. When $\beta \neq 0$, one defines a unique scale parameter $C = \frac{1}{2}(C_+ + C_-)$, which together with β allows one to describe the behavior at $x \to \pm\infty$. The completely antisymmetric case $\beta = +1$ (resp. -1) corresponds to the maximum asymmetry.

For $0 < \mu < 1$ and $\beta = \pm 1$, the random variables take only positive (resp. negative) values.

For $1 < \mu < 2$ and $\beta = +1$, the Lévy law is a power law for $x \to +\infty$ but goes to zero for $x \to -\infty$ as $P(x) \sim \exp(-|x|^{\mu/\mu-1})$. This decay is faster than the Gaussian law. The symmetric situation is found for $\beta = -1$.

An important consequence of (A.2) is that the variance of a Lévy law is infinite as the p.d.f. does not decay sufficiently rapidly at $|x| \to \infty$. When $\mu \leq 1$, the Lévy law decays so slowly that even the mean and the average

of the absolute value of the spread diverge. The median and the most probable value still exist and coincide, for symmetric p.d.f. ($\beta = 0$), with the center m. The characteristic scales of the fluctuations are determined by the scale parameter C, i.e. they are of the order of $C^{1/\mu}$.

There are no simple analytic expression of the symmetric Lévy stable laws $L_\mu(x)$, except for a few special cases. The best known is $\mu = 1$, called the Cauchy (or Lorentz) law,

$$L_1(x) = \frac{1}{x^2 + \pi^2} \quad \text{for} \quad -\infty < x < +\infty. \tag{A.5}$$

The Lévy law for $\mu = \frac{1}{2}$ is [19]

$$L_{1/2}(x) = \frac{2}{\sqrt{\pi}} \frac{\exp(-1/2x)}{(2x)^{3/2}} \quad \text{for} \quad x > 0. \tag{A.6}$$

This p.d.f. $L_{1/2}(x)$ gives the distribution of first returns to origin of an unbiased random walk.

Lévy laws are fully characterized by the expression of their characteristic functions:

$$\hat{L}_\mu(k) = \exp(-a_\mu |k|^\mu), \tag{A.7}$$

where a_μ is a constant proportional to the scale parameter C:

$$a_\mu = \frac{\pi C}{\mu^2 \Gamma(\mu - 1) \sin(\pi\mu/2)} \quad \text{for} \quad 1 < \mu < 2. \tag{A.8}$$

A similar expression holds for $0 < \mu < 1$, while $\mu = 1$ and 2 requires a special form (see [10] for full details). For $\beta \neq 0$, we have

$$\hat{L}_\mu^\beta(k) = \exp\left[-a_\mu |k|^\mu \left(1 + \mathrm{i}\beta \tan\left(\frac{\mu\pi}{2}\right) \frac{k}{|k|}\right)\right] \quad \text{for} \quad \mu \neq 1. \tag{A.9}$$

For $\mu = 1$, $\tan(\mu\pi/2)$ is replaced by $(2/\pi) \ln |k|$.

Appendix B. Proof of the optimality of the solution of (44)

To prove the optimality of the solution $\mathbf{K}^{\mathrm{o}}$ solving (44), it is sufficient to consider only one of the system of N equations for a single and fixed index i. We thus drop the index i and define the matrix $\boldsymbol{\Omega}^{\mathrm{f,o}} \in \mathbb{R}^{N \times N}$ and the vector $\boldsymbol{\kappa} \in \mathbb{R}^N$ as follows:

$$\Omega_{pq}^{\mathrm{f}} = \begin{cases} |G_{ip}^{\mathrm{f}} - \sum_{m=1}^N K_{im}^{\mathrm{o}} G_{mp}^{\mathrm{f}}|^{\mu-2} C_p^{\mathrm{f}} & \text{for} \quad p = q, \\ 0 & \text{for} \quad p \neq q, \end{cases}$$

$$\Omega_{pq}^{\mathrm{o}} = \begin{cases} |\sum_{m=1}^N K_{im}^{\mathrm{o}} G_{mp}^{\mathrm{o}}|^{\mu-2} C_p^{\mathrm{o}} & \text{for} \quad p = q, \\ 0 & \text{for} \quad p \neq q, \end{cases}$$

$$\kappa_q = K_{iq}^{\mathrm{o}}. \tag{B.1}$$

Expression (45) can then be written as

$$\frac{\partial^2}{\partial \boldsymbol{\kappa}^2} \hat{B}_{ii} = \mu(\mu - 1)[\mathbf{G}^{\mathrm{f}} \boldsymbol{\Omega}^{\mathrm{f}} (\mathbf{G}^{\mathrm{f}})^{\mathrm{T}} + \mathbf{G}^{\mathrm{o}} \boldsymbol{\Omega}^{\mathrm{o}} (\mathbf{G}^{\mathrm{o}})^{\mathrm{T}}]. \tag{B.2}$$

It is clear in this form that the Hessian matrix is positive definite for $\mu > 1$ because the linear sum of symmetric positive definite matrices results in a positive definite matrix.

There are two issues associated with this formulation for the optimal $\boldsymbol{\kappa}$ (i.e., $\mathbf{K}^{\mathrm{o}}$): (1) the possible existence of singular terms for $\mu < 2$; and (2) the solvability of the nonlinear system. To address these issues, we rewrite (44) as

$$\frac{\partial}{\partial \boldsymbol{\kappa}} \operatorname{trace} \hat{B}_{ii} = \mathbf{f}(\boldsymbol{\kappa}, \mu) = 0, \tag{B.3}$$

and seek for a solution branch $\boldsymbol{\kappa}(\mu)$ of (B.3) using implicit function theory as μ varies.

At $\mu = 2$, the system is linear in $\boldsymbol{\kappa}$ and a unique solution $\boldsymbol{\kappa}(\mu = 2)$ can be obtained analytically (this is nothing but the standard linear least-variance estimation). For $1 < \mu < 2$, $\mathbf{f}(\boldsymbol{\kappa}, \mu)$ is bounded. Its derivative with respect to $\boldsymbol{\kappa}$, $(\partial/\partial\boldsymbol{\kappa})\mathbf{f}(\boldsymbol{\kappa}, \mu)$, is given by (B.2) which is always positive definite. It can be singular and diverge to $+\infty$ due to absolute value terms behaving like $\lim_{x\to 0} |x|^{\mu-2} = \infty$ as can be seen in Eq. (B.1). The derivative of $\mathbf{f}(\boldsymbol{\kappa}, \mu)$ with respect to μ is

$$\begin{aligned}\frac{\partial}{\partial \mu}\mathbf{f}(\boldsymbol{\kappa}, \mu) = \frac{\partial^2}{\partial \boldsymbol{\kappa}\,\partial \mu}\hat{B}_{ii} = \mu \sum_{p=1}^{N} \Bigg[&- \left(\log\left|\left(G^{\mathrm{f}}_{ip} - \sum_{m=1}^{N} K^{\mathrm{o}}_{im} G^{\mathrm{f}}_{mp}\right)\right|\right)\left(G^{\mathrm{f}}_{ip} - \sum_{m=1}^{N} K^{\mathrm{o}}_{im} G^{\mathrm{f}}_{mp}\right)^{[\mu-1]} G^{\mathrm{f}}_{jp} C^{\mathrm{f}}_{p} \\ &+ \left(\log\left|\left(\sum_{m=1}^{N} K^{\mathrm{o}}_{im} G^{\mathrm{o}}_{mp}\right)\right|\right)\left(\sum_{m=1}^{N} K^{\mathrm{o}}_{im} G^{\mathrm{o}}_{mp}\right)^{[\mu-1]} G^{\mathrm{o}}_{jp} C^{\mathrm{o}}_{p} \Bigg].\end{aligned} \tag{B.4}$$

This derivative can also be singular due to the absolute value terms behaving like $\lim_{x\to 0} x^{[\mu-1]} \log|x| = \pm\infty$. Note that $(\partial/\partial\boldsymbol{\kappa})\mathbf{f}(\boldsymbol{\kappa}, \mu)$ and $(\partial/\partial\mu)\mathbf{f}(\boldsymbol{\kappa}, \mu)$ become singular simultaneously, but the former is more singular than the latter because $\lim_{x\to 0}|x|^{\mu-2}/x^{[\mu-1]} \log|x| = \pm\infty$. Implicit function theory can therefore be applied to guarantee that a unique solution branch $\boldsymbol{\kappa}(\mu)$ exists for $1 < \mu < 2$, starting from the analytical solution at $\mu = 2$. Indeed, if there exist other solution branches, then there must be at least one bifurcation as μ varies because the solution at $\mu = 2$ is globally unique due to linearity. The fact that $(\partial/\partial\boldsymbol{\kappa})\mathbf{f}(\boldsymbol{\kappa}, \mu)$ is a positive definite matrix globally (though it can be singular), however, guarantees that there is no bifurcation. Accordingly, the solution branch from $\mu = 2$ provides the unique solution of the system as μ varies.

The solution on the branch can be obtained numerically, either by directly solving for the N nonlinear equations for each μ or by following the branch using the pseudo arc-length continuation method [14].

References

[1] H. Ahn, R.E. Feldman, Optimal filtering of a Gaussian signal in the presence of Lévy noise, SIAM J. Appl. Math. 60 (1999) 359–369.
[2] F.A. Aliev, L. Ozbek, Evaluation of convergence rate in the central limit theorem for the Kalman filter, IEEE Trans. Automatic Control 44 (1999) 1905–1909.
[3] J.-P. Bouchaud, D. Sornette, C. Walter, J.-P. Aguilar, Taming large events: optimal portfolio theory for strongly fluctuating assets, Int. J. Theoret. Appl. Finance 1 (1998) 25–41.
[4] J.L. Chambers, C.L. Mallows, B.W. Stuck, A method for simulating stable random variables, J. Am. Statist. Assoc. 71 (1976) 340–344.
[5] C. de Calan, J.-M. Luck, T.M. Nieuwenhuizen, D. Petritis, On the distribution of a random variable occurring in 1D disordered systems, J. Phys. A 18 (1985) 501–523.
[6] R.A. Daley, Atmospheric Data Analysis, Cambridge University Press, Cambridge, 1991, 457 pp.
[7] W. Feller, An Introduction to Probability Theory and its Applications, Vol. 2, Wiley, New York, 1966–1971.
[8] M. Ghil, P. Malanotte-Rizzoli, Data assimilation in meteorology and oceanography, Adv. Geophys. 33 (1991) 141–266.
[9] M. Ghil, S. Cohn, J. Travantzis, K. Bube, E. Isaacson, Application of estimation theory to numerical weather prediction, in: L. Bengtsson, M. Ghil, E. Källén (Eds.), Dynamic Meteorology, Springer, New York, 1981.
[10] B.V. Gnedenko, A.N. Kolmogorov, Limit Distributions for Sum of Independent Random Variables, Addison-Wesley, Reading, MA, 1954.
[11] C.M. Goldie, Implicit renewal theory and tails of solutions of random equations, Ann. Appl. Probab. 1 (1991) 126–166.

[12] K. Ide, P. Courtier, M. Ghil, A. Lorenc, Unified notation for data assimilation: operational, sequential and variational, J. Meteorol. Soc. Jpn. 75 (1997) 181–189.
[13] N.L. Johnson, S. Kotz, N. Balakrishnan, Continuous Univariate Distributions, Vol. 2, Wiley, New York, 1994.
[14] H.B. Keller, Numerical solution of bifurcation and nonlinear eigenvalue problems, in: P.H. Rabinowitz (Ed.), Application of Bifurcation Theory, Academic Press, New York, 1977.
[15] H. Kesten, Random difference equations and renewal theory for products of random matrices, Acta Math. 131 (1973) 207–248.
[16] A. Le Breton, M. Musiela, A generalization of the Kalman filter to models with infinite variance, Stochastic Process. Appl. 47 (1993) 75–94.
[17] J.H. McCulloch, Numerical Approximation of the Symmetric Stable Distributions and Densities, Ohio State University Economics Department, 1994. http://economics.sbs.ohio-state.edu/jhm/jhm.html.
[18] R.N. Miller, M. Ghil, F. Gauthiez, Advanced data assimilation in strongly nonlinear dynamical systems, J. Atmos. Sci. 51 (1994) 1037–1056.
[19] E.W. Montroll, B.J. West, On an enriched collection of stochastic processes, in: E.W. Montroll, J.L. Lebowitz (Eds.), Fluctuation Phenomena, Elsevier, Amsterdam, 1979.
[20] S.P. Nishenko, C.C. Barton, Scaling laws for natural disasters: application of fractal statistics to life and economic loss data (abstract), Geol. Soc. Am. (Abstracts with Programs) 25 (1993) 412.
[21] G. Samorodnitsky, M.S. Taqqu, Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance, Chapman & Hall, New York, 1994.
[22] D. Sornette, Linear stochastic dynamics with nonlinear fractal properties, Physica A 250 (1998) 295–314.
[23] D. Sornette, R. Cont, Convergent multiplicative processes repelled from zero: power laws and truncated power laws, J. Phys. I. France 7 (1997) 431–444.
[24] D. Sornette, Critical Phenomena in Natural Sciences. Chaos, Fractals, Self-organization and Disorder: Concepts and Tools, Springer Series in Synergetics, Springer, Heidelberg, 2000.
[25] J.C. Spall, K.D. Wall, Asymptotic distribution theory for the Kalman filter state estimator, Commun. Statist. Theory Meth. 13 (1984) 1981–2003.
[26] D. Zajdenweber, Business interruption insurance, a risky business — a study on some Paretian risk phenomena, Fractals 3 (1995) 601–608.
[27] D. Zajdenweber, Extreme values in business interruption insurance, J. Risk Insurance 63 (1996) 95–110.