

## ARTICLE

## Reappraisal of predictive models for resuspension

Shigeru Kumazawa\*

Formerly, Japan Nuclear Energy Safety Organization, 4-1-28 Toranomon, Minato-Ku, Tokyo, 105-0001, Japan

A statistical reappraisal of models is presented for the prediction of resuspension with the assessment of uncertainty, using the historical data reported by Maxwell and Anspaugh (2011). A model, conveniently called “hybrid scale model”, was applied to unify the power law and the exponential function, which are often used in conventional predictive models of resuspension factor ( $S_f$ ), and also to reflect the bounded data of  $S_f$ . The hybrid scale model has been proven to be well fitted to the data of resuspension factor with the improved assessment of uncertainty as a linearized model between observation time and resuspension factor.

**Keywords:** accident; nuclear; resuspension factor; predictive model; uncertainty; hybrid log-normal; hybrid scale; nonlinear regression analysis

## 1. Introduction

After the Fukushima Daiichi nuclear accident, the resuspension factor ( $S_f$ ) has been quite often used to calculate the inhalation dose from the resuspended nuclides of surface contamination. To attain the realistic calculation, we must consider the strong dependence on time in the resuspension factor. Maxwell and Anspaugh [1] comprehensively reviewed the predictive models and revised their former model of double exponential for prediction of resuspension (2002) [1], to be a more accurate predictor of  $S_f$  at early times (observation time  $t < \sim 100$  d) and a practical model of analytical solutions.

Most of the models are given as a function of  $\ln(S_f) = \alpha + \beta g(t)$  where  $\alpha, \beta$  are a constant, and  $g(t) = \ln(t)$  or  $t$  of time. Maxwell and Anspaugh showed a good fit of  $\ln(S_f)$  vs.  $t$  for  $t < 100$  d but  $\ln(S_f)$  vs.  $\ln(t)$  for  $t > 100$  d. However the hybrid function  $hyb(x) = x + \ln(x)$  [2] is feasible for the predictive model over the whole period, putting  $g(t) = hyb(\tau t)$  with a positive constant,  $\tau (d^{-1})$ .

The range of  $S_f$  is reported from  $10^{-2}$  to  $10^{-10} \text{ m}^{-1}$  [3] or  $10^{-5}$  to  $10^{-9} \text{ m}^{-1}$  [1]. Thus the function  $\ln[(S_f - a)/(b - S_f)]$  or  $hyb[v(S_f - a)/(b - S_f)]$  is feasible for both bounded data  $a < S_f < b$  where  $v$  is a constant. The value of lower boundary,  $a$ , should reflect the argument that  $S_f$  cannot have been greater than  $7 \times 10^{-11}$  or  $2 \times 10^{-10} \text{ m}^{-1}$ , 15 or 18 years after deposition, respectively [4].

In this paper models are proposed for the prediction of resuspension to attain a linear relationship between time  $t$  and  $S_f$ , both transformed by some functions, to the data obtained from the reference [1]. The uncertainty is evaluated to provide the prediction intervals of  $S_f$  based on the normality of residual errors.

## 2. Method

## 2.1. Distribution analysis on homogenous sampling

The statistical homogeneity of time  $t$  and  $S_f$  within the dataset is important to properly estimate the time function of resuspension factor against the distortion due to not evenly sampling in time and conditions.

The distribution of time  $t$  is analyzed by the hybrid lognormal (HLN) distribution that  $hyb(\tau t)$  is normally distributed with the mean  $\mu$  and the variance  $\sigma^2$ , where  $\tau (d^{-1})$  is a constant, adjusting between the power law and the exponential function in observation times.

Parameters  $\tau$ ,  $\mu$ , and  $\sigma$  are estimated to minimize the sum of squared errors (SSE)  $\sum \varepsilon_i^2$  or to maximize the coefficient of determination ( $R^2$ ) to sorted data  $\{t_i | i = 1, n\}$  in the equation,  $hyb(\tau t_i) = \mu + \sigma z_i + \varepsilon_i$ , where the  $i$ th smallest  $z_i$  is given by the normal distribution function  $\Phi(z_i) \approx (i - 3/8) / (n + 1/4)$  [5] and  $n$  is the number of data.

The distribution of  $S_f$  is analyzed by the extreme value type I distribution function,  $F(u) = \exp[-\exp(-u)]$ , putting the equation,  $hyb[v(S_f - a)] = \mu + \eta u_i + \varepsilon_i$ , conveniently called “hybrid Freshet 4-parameter (HF4) distribution,” where  $u_i$  is given by  $F(u_i) \approx (i - 0.3175)/(n + 0.365)$  but  $F(u_i) = 1 - 0.5^{1/n}$ ,  $F(u_n) = 0.5^{1/n}$  [6]. Parameters  $v$ ,  $a$ ,  $\mu$ , and  $\eta$  are estimated by  $\sum \varepsilon_i^2 \rightarrow \min$  or  $R^2 \rightarrow \max$  on data  $\{S_{fi}\}$ . HF4 is preferable based on  $AIC = n \ln(SSE/n) + 2\theta$  ( $\theta$ : the number of parameters), due to the extremely large variation of  $S_f$  over the whole period (0.04 ~ 2000 d).

However, we also used the lognormal distribution to data of  $S_f$  in each of divided observation periods. Distributions of  $S_{fj}$  among the  $j$ th period of 22 time intervals properly selected were analyzed with the time function of  $\mu$  and  $\sigma$  to obtain the lognormal (LN) distribution,  $\ln(S_{fj}) \sim N(\mu_j, \sigma_j^2)$ , ( $j = 1 \sim 22$ ).

\*Corresponding author. Email: shuen85@jcom.home.ne.jp

2.2. Relationship analysis between time  $t$  and  $S_f$

The best linearity between  $x$  and  $y$  is essential to assess the uncertainty of fitting the model to data. The linear relationship is well known about data plotted on normal, two types of semi-log, and log-log section papers, but data  $\{t, S_f\}$  are often not fitted with a straight line on the four kinds of popular section papers.

The hybrid function  $hyb(x)$  affords unification to the four kinds of popular section papers because the hybrid scale is continuously connected between the log scale and the linear scale shown as in **Figure 1** and both axes in hybrid scale provides nine types of section papers.

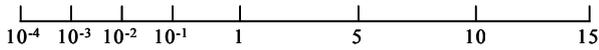


Figure 1. An example of hybrid scales graduated in  $x$  on the linear scale of  $y$  via the function  $y = hyb(x)$ .

Table 1. Nine types of linear relationships on hybrid scale.

Linear-Log $y = \alpha + \beta \ln(x)$	Linear-Hybrid $y = \alpha + \beta hyb(x)$	Linear-Linear $y = \alpha + \beta x$
Hybrid-Log $hyb(y) = \alpha + \beta \ln(x)$	Hybrid-Hybrid $hyb(y) = \alpha + \beta hyb(x)$	Hybrid-Linear $hyb(y) = \alpha + \beta x$
Log-Log $\ln(y) = \alpha + \beta \ln(x)$	Log-Hybrid $\ln(y) = \alpha + \beta hyb(x)$	Log-linear $\ln(y) = \alpha + \beta x$

The hybrid scale (HS) model is defined as nine types of linear relationships on the section papers derived from applying the hybrid scale to both axes of  $x$  and  $y$  (see **Table 1**). For the prediction of resuspension the HS model is applied variously, e.g.,  $x = \tau t$  and  $y = \nu S_f$ .

The relationships are examined as follows;

Preliminary analysis of piecewise data by period

$$hyb[\nu \cdot \{-\mu_j - a\} / \{b - (-\mu_j)\}] = \alpha + \beta hyb(\tau t_j) + \epsilon_j, \quad (1)$$

$$\sigma_j = \alpha + \beta hyb(\tau t_j) + \epsilon_j \quad (2)$$

$$hyb[\nu \cdot (S_{fjgm} - a)] = \alpha + \beta hyb(\tau t_j) + \epsilon_j, \quad (3)$$

where  $S_{fjgm}$  is the geometric mean  $\exp(u_j)$  of  $S_{fj}$ , and  $\tau, \nu, a, b, \alpha$  and  $\beta$  are model parameters to be estimated by  $\Sigma \epsilon_j^2 \rightarrow \min$  or  $R^2 \rightarrow \min$  for data  $\{t_j, \mu_j\}, \{t_j, \sigma_j\}$ , or  $\{t_j, S_{fjgm}\}$ , ( $j = 1$  to  $22$ ), over 22 divided periods;

Main analysis of all data over whole period

$$\ln(S_{fi}), hyb[\nu \cdot (S_{fi} - a)] \text{ or } hyb[\nu \cdot (S_{fi} - a) / (b - S_{fi})] = \alpha + \beta hyb(\tau t_i) + \epsilon_i, \quad (4)$$

where  $\tau, \nu, a, b, \alpha$  and  $\beta$  are model parameters to be estimated for data  $\{t_i, S_{fi}\}$ , ( $i = 1$  to  $295$ ), to minimize  $\Sigma \epsilon_i^2$  or to maximize  $R^2$ , via selecting the smaller AIC model.

2.3. Uncertainty analysis on estimated functions  $S_f(t)$

Attaining the best linearity so that the residual errors should be normally distributed as much as possible, we can use the prediction intervals ( $PI$ ) for a future value  $Y$ , given explanatory level  $x = hyb(\tau t)$ , assumed the known values of  $\tau, \nu, a$  and  $b$  calculated in advance, as follows:

$$\hat{y} \pm t_{\alpha/2, s} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \quad (5)$$

where  $\hat{y}$  is the estimate of the response variable (e.g.  $y_i = hyb[\nu \cdot (S_{fi} - a)]$  or  $hyb[\nu \cdot (S_{fi} - a) / (b - S_{fi})]$ ),  $t_{\alpha/2}$  is lower and upper prediction levels of the student's  $t$  distribution with the degree of freedom  $n - k - 1$  for  $k$  parameters except for intercept,  $s$  is the residual standard error,  $\bar{x}$  is the mean of explanatory variable  $x_i = hyb(\tau t_i)$ .

Parameters  $(\tau, \nu, a, b)$  estimated nonlinearly should be considered as factors of influencing the uncertainty. This paper, however, considers their influence on the degree of freedom only because of avoiding the complexity.

To compare with the estimation of uncertainty by Maxwell and Anspaugh [1],  $\alpha/2 = 15.87\%$  is calculated, which corresponds to one sigma of the normal distribution, while  $\alpha/2 = 2.5\%$  is widely used.

3. Results

3.1. Distribution analysis on homogenous sampling

Using 295 data read by eyeball as well as possible from the several figures, on which data ( $n > 300$ ) from 0.04 to 2000 d were plotted variously by Maxwell and Anspaugh [1], we identified the probability distribution to fit well to all data of observation time  $t$  and  $S_f$ .

**Figure 2** shows that observation time  $t$  is hybrid lognormally distributed, where lognormally distributed at early times ( $t < 100$  d), and normally distributed for time greater than 100 d. **Figure 3** shows that  $S_f$  follows the HF4 distribution, where the Frechet 3-parameter distribution for  $S_f < \sim 10^{-5} \text{ m}^{-1}$  but the extreme value type I distribution for  $S_f > \sim 10^{-5} \text{ m}^{-1}$ .

Although these 295 data are probably not random, we could find a statistical homogeneity in terms of observation time ( $t$ ) and resuspension factor ( $S_f$ ).

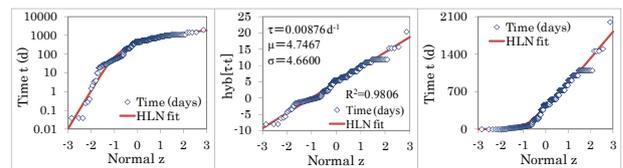


Figure 2. Normal probability plot of data  $t$  (d). From left, log plot, hybrid plot and linear plot.

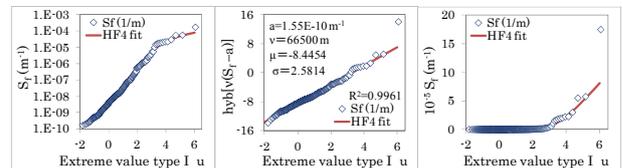


Figure 3. Extreme value type I probability plot of data  $S_f$  ( $1/m$ ). From left, log plot, hybrid plot and linear plot.

3.2. Relationship analysis between time  $t$  and  $S_f$

(a) Preliminary analysis on piecewise data by period

Resuspension factors  $S_{fj}$  during the  $j$ th period were proved to be expressed as  $\ln(S_{fj}) \sim N(\mu_j, \sigma_j^2)$ , ( $j = 1$  to  $22$ ). The degree of  $R_j^2$  for  $j$ th period is more than 0.93, except two periods where  $R_j^2 = 0.70, 0.90$ .

In **Figure 4** the parameter  $\mu_j$  of LN decreases by time after deposition along the straight line (left panel) or the

curve (left panel) of the HS model given as follows:

$$\text{hyb} \left[ 0.193 \frac{-\mu - 9.295}{21.76 + \mu} \right] = -0.975 + 0.453 \text{hyb}[0.005t]. \quad (6)$$

In **Figure 5** the parameter  $\sigma_j$  of LN is less dependent on time after deposition despite the large variation but can be fitted as  $\sigma = 1.473 - 0.137 \text{hyb}[0.005t]$ , which is similar to  $1.44 = \ln(\delta)$  shown as the uncertainty of  $S_f(t)$  in the reference [1].

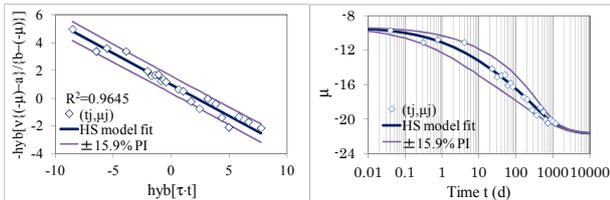


Figure 4. Hybrid plot of  $\mu_j$  vs. geometric mean time in the  $j$ th period. Left: hybrid-hybrid plot, Right: semi-log plot.

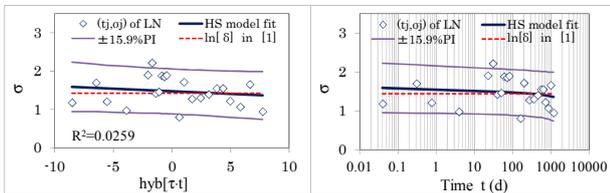


Figure 5. Hybrid plot of  $\sigma_j$  vs. geometric mean time, to the  $j$ th period. Left: semi-hybrid plot, Right: semi-log plot.

According to the lognormality of  $S_{fj}$  within each of periods selected, we can obtain the time function  $S_f(t)$ , shown on the left panel and the right panel of **Figure 6**, respectively, by substituting  $\ln(S_{fjgm})$  for  $\mu_j$  in Eq. (1) or by fitting Eq. (3) to data  $S_{fjgm} = \exp(\mu_j)$ . Both results of  $S_f(t)$  are similar for  $t < 1500$  d. The curve of  $S_f(t)$  on the right panel is given by  $\text{hyb}[66500(S_f - 5.25 \times 10^{-10})] = -8.460 - 1.373 \text{hyb}[0.001 t]$ , where  $R^2 = 0.9722$ .

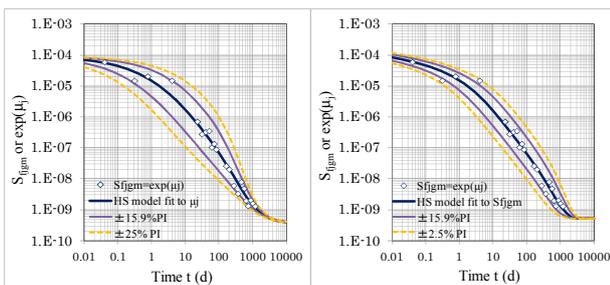


Figure 6. Curves of  $S_f(t)$  calculated from LN fit by period. Left:  $\mu_j = \ln(S_{fjgm})$  in Eq.(1), Right: fit Eq.(3) to  $S_{fjgm} = \exp(\mu_j)$ .

(b) Main analysis of all data over the whole period

For the improved assessment of uncertainty, we should consider the individual variation of 295 data  $\{t_i, S_{fi}\}$ :  $0.04 \leq t \leq 2000$  d and  $1.7 \times 10^{-10} \leq S_f \leq 1.75 \times 10^{-4} \text{ m}^{-1}$ .

First,  $\ln(S_f) \propto \tau t + \ln(\tau t)$  or  $\ln(S_f) \propto \text{hyb}(\tau t)$  is derived as Eq. (4), corresponding to the reference [1] " $\ln(S_f) \propto \ln(t)$  at early times ( $< 100$  d) and  $\ln(S_f) \propto t$  at long times ( $> 100$  d). In fact, the hybrid lognormality of observation times in  $\text{hyb}(\tau t)$  is shown in Figure 2.

Second, it is natural to consider some upper and lower

constraints on  $S_f$ ,  $a < S_f < b$ :  $10^{-5}$  to  $10^{-9} \text{ m}^{-1}$  [1],  $10^{-2}$  to  $10^{-10} \text{ m}^{-1}$  [3],  $\leq 7 \times 10^{-11}$  or  $\leq 2 \times 10^{-10} \text{ m}^{-1}$  for 5475 d or 6570 d [4], respectively, etc. Available data of  $S_f$  for times ( $> 1000$  d or  $< 10$  d) is, however, a few. Thus the lower limit  $a$  is selected to be  $1.5 \times 10^{-10} \text{ m}^{-1}$  that is less than the minimum  $1.7 \times 10^{-10} \text{ m}^{-1}$  of 295 data available. The upper limit  $b$ , however, can be neglected for the moderate constraint as  $\text{hyb}[v \cdot (S_{fi} - a)]$  in Eq. (4) due to the characteristics of hybrid scale, while  $b$  should be needed for the strong constraint as  $\text{hyb}[v \cdot (S_{fi} - a) / (b - S_{fi})]$ .

**Figure 7** shows that the HS model of Eq. (7) is well fit to 295 data, scattered along a straight line on the left panel, except two points ( $t=0.04$  d,  $S_f=1.75 \times 10^{-4} \text{ m}^{-1}$ ,  $t=56$  d,  $S_f=5 \times 10^{-10} \text{ m}^{-1}$ ), where  $R^2$  is 0.7314:

$$\text{hyb}[66500(S_f - 1.5 \times 10^{-10})] = -8.36 - 1.40 \text{hyb}[0.001t]. \quad (7)$$

The curve on the right panel in Figure 7 is very similar to both curves in Figure 6 for  $< 1000$  d but the 95% prediction band ( $\pm 2.5\%$  PI) is wider than those in Figure 6 for  $t > 1$  d. As the prediction interval includes the contribution due to individual variation of data, the 295 data results in the larger uncertainty of random errors.

Maxwell and Anspaugh [1] selected the upper limit  $10^{-5} \text{ m}^{-1}$ . If such a strong upper constraint is possible,  $\text{hyb}[v \cdot (S_{fi} - a) / (b - S_{fi})]$  in Eq. (4) will be applicable. In fact it works. We, however, need more data at quite early times to confirm the preference of  $\text{hyb}[v \cdot (S_{fi} - a) / (b - S_{fi})]$ .

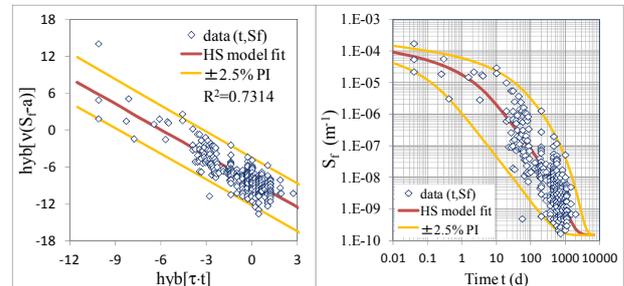


Figure 7. Hybrid plot of  $S_{fi}$  vs.  $t_i$  ( $i=1, 295$ ). Left: hybrid-hybrid plot, Right: log-log plot.

3.3. Uncertainty analysis on estimated functions  $S_f(t)$

The uncertainty analysis of  $S_f$  should be based on the significant factors as discussed in WASH-1400 [4], IAEA-TECDOC-647 [3], etc. This paper, however, focused on the statistical aspect of it practically.

According to the classical equation for simple linear regression, Eq. (5) provides the adequate PI for a future response  $y = \text{hyb}[v \cdot (S_{fi} - a)]$ , given an explanatory level  $x = \text{hyb}(\tau t)$ , where  $\tau$  is assumed known as parameters estimated in advance.

In Figure 4 or 5, the 68% confidence band (CB) was constructed as the  $\pm 15.9\%$  PI for the plots linearized on the left panel, and the  $\pm 15.9\%$  PI on the right panel were simply mapped from the  $\pm 15.9\%$  PI on the left panel, to avoid the complexity. The 15.9% PI is due to the small size of data  $\{t_j, S_{fjgm}\}$ .

In Figure 6, the 68 % CB of  $S_{fjgm}$  on the left panel was

simply mapped from one on the right panel of Figure 4 as  $\exp(\pm 15.9\% PI)$  but that on the right panel was simply mapped from the 68 % CB of the linearized plots by fitting Eq. (3) to data  $S_{fjgm} = \exp(\mu_j)$ . Similarly the 95% CBs were shown as the  $\pm 2.5\%$  PI on both panels. Despite the different methods to obtain  $S_f(t)$  via  $S_{fjgm}$ , the confidence band on the right panel is narrower than one on the left panel. In Figure 7, the 95% CB of 295 data  $\{t_i, S_{fi}\}$  on right panel was constructed in the same way by simply mapping the  $\pm 2.5\%$  PI for the plots linearized on the left panel.

The uncertainty of  $S_f(t)$  based on data  $\{t_{jgm}, S_{fjgm}\}$  is explicitly smaller than that based on all data  $\{t_i, S_{fi}\}$ . The former uncertainty (piecewise data) is comparable in magnitude to the uncertainty in the reference [1] as the  $\pm 15.9\%$  PI (nearly  $\pm$ one sigma) for  $\sim 100$  to  $\sim 200$  d but it is roughly 1/3 at 100 d, 1/4 at 200 d and 1/10 at 500 d of the latter uncertainty (all data). Thus it is important to improve the uncertainty assessment.

Historically various models [1, 3, 4] have been proposed for  $S_f(t)$  as shown on the left panel of **Figure 8**. Among models the latest one is the improved version of the double exponential model [1] (right panel, Figure 8), which is within the painted area of data measured in Europe after the Chernobyl accident on the left panel.

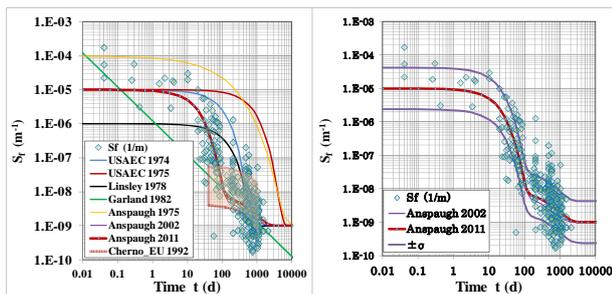


Figure 8. Conventional models of  $S_f(t)$ . Left: various models, Right: the improved model of Maxwell and Anspaugh [1].

The double exponential model has been derived based on the homogenous data in time divided into  $t < 100$  d and  $t > 100$  d. The statistical homogeneity of time  $t$  and  $S_f$ , however, has been proved here by the HLN and HF4 distributions, and then we can estimate a predictive models of  $S_f(t)$  on all 295 data without piecewise analysis to improve the assessment of uncertainty.

In general as the uncertainty assessment strongly depends on the judgment of experts who know the details of resuspension factor and regulatory practice, it is important to consult the original publications to ensure that the way the parameter values were originally obtained is compatible with they are to be used in assessment calculations, as recommended by IAEA [7].

The value of  $S_f = 10^{-6} \text{ m}^{-1}$  is widely used because it is considered to be the upper bound assuming weathered deposition [8]. The latest IAEA report [7] summarizes “in the first days and months after deposition, the value of  $S_f$  generally ranges between  $10^{-5} \text{ m}^{-1}$  in residential areas, on sites undergoing cleanup operations and on

arid sites, and  $10^{-6} \text{ m}^{-1}$  on rural sites.”

Therefore the reappraisal of  $S_f(t)$  is expected for data newly available from the recent nuclear accident. Then the hybrid scale model can be also used in various ways of data analysis as shown here.

#### 4. Conclusion

In this paper, a linearized model was presented for predicting the resuspension factor ( $S_f$ ) by applying the hybrid scale (HS) model [2] to the historical dataset [1]. The HS model with 5 parameters was found to fit well to the data of  $S_f$  over the whole observation times due to the smaller AIC (the model selection criteria) of 325.22 for this model, while it is 383.93 as estimated for the latest conventional model with 5 parameters [1]. The HS model demonstrated to provide the prediction intervals for a future value of  $S_f$  over the observation period so that those should reflect the dataset more realistically.

The HS model can be applicable for analyzing additional data to be obtained in studies on the Fukushima Daiichi nuclear accident.

#### References

- [1] R.M. Maxwell and L.R. Anspaugh, An improved model for prediction of resuspension, *Health. Phys.* 101 (2011), pp. 722-730.
- [2] S. Kumazawa, A hybrid-scale theory to be applied in health risk assessment, *Comparative Evaluation of Environmental Toxicant: Health Effects of Environmental Toxicants Derived from Advanced Technologies*, Edited by J. Inaba and Y. Nakamura, Kodansha Scientific Ltd., Tokyo, (1998), pp. 177-190, ISBN 4-906464-06-8.
- [3] USAEC, *Reactor Safety Study - An Assessment of Accident Risks in US Commercial Nuclear Power Plants* -, WASH-1400, Appendix VI, (1992), pp. E-9-E-14.
- [4] J.A. Garland, N.J. Pattenden and K. Playford, Chapter 2. Resuspension Following Chernobyl, Modelling of resuspension seasonality and losses during food processing, *First Report of the VAMP Terrestrial Working Group*, IAEA-TECDOC-647, (1992), pp. 9-33.
- [5] G. Blom, *Statistical Estimates and Transformed Beta Variables*, New York, John Wiley & Sons, (1958).
- [6] NIST, *Engineering Statistics Handbook* ([http://www.itl.nist.gov/div898/handbook/eda/section3/pr\\_obplot.htm](http://www.itl.nist.gov/div898/handbook/eda/section3/pr_obplot.htm)).
- [7] IAEA, *Handbook of Parameter Values for the Prediction of Radionuclide Transfer in Terrestrial and Freshwater Environments*, Technical Reports Series No.472 (2010).
- [8] IAEA, *Generic Procedures for Assessment and Response during a Radiological Emergency*, IAEA- TECDOC-1162, (2000), pp.96-101.