

Simple Statistical System for
Assimilation of Models and Monitoring data

Air4EU program tool SAM version 1.0

Sam-Erik Walker NILU
Air4EU Expert Workshop
30 June 2006

Air4EU DA tool: Background

- Program tool developed to enable users to perform **simple data assimilation** as part of the Air4EU project
- Currently written as a Windows based console application (executable file **sam.exe**)
- Later version may be made as a .NET Web application with graphics integrated

Air4EU DA tool: Philosophy

- The tool is based on reading in one model concentration for each of the spatial scales **Local**, **Urban** and **Regional**:

C_L , C_U and C_R

together with an Observed concentration:

C_O

- The concentrations can be given in any unit, but the unit used and the averaging time must be the same for all data
- We assume that the concentrations are representative for a given receptor point within our area of interest

Air4EU DA tool: Philosophy

- To compare model concentrations with the observed concentration we must first calculate a **Total Model concentration**:

$$C_M = C_L + C_U + C_R$$

- The **Total Model concentration** C_M will be combined with the **Observed concentration** C_O to form a new **Assimilated concentration** C_A
- The Assimilated concentration C_A represents a best estimate of what the True concentration C_T is given the Total Model and Observed concentrations

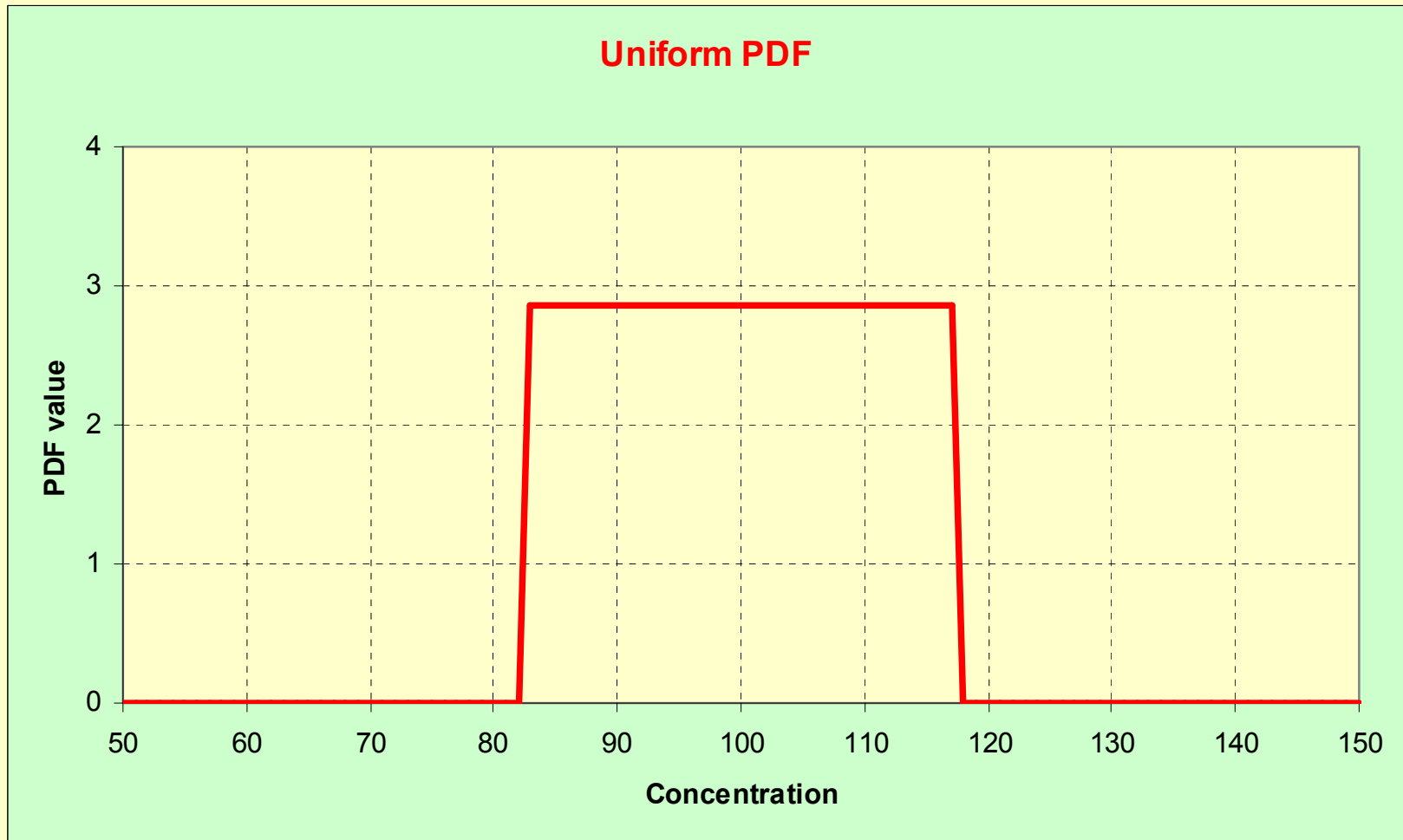
Air4EU DA tool: Philosophy

- For the data assimilation procedure to be meaningful we must specify the **uncertainty** associated with each concentration value C_L , C_U , C_R and C_O
- Uncertainties are best described using the language of statistics, i.e., by using **probabilities**
- The uncertainty associated with each concentration is defined as a **probability distribution** (or PDF) $p(c)$ associated with each concentration
- The PDF is uniquely fixed by defining e.g., its **mode** (most probable value) (C) and its **standard deviation**

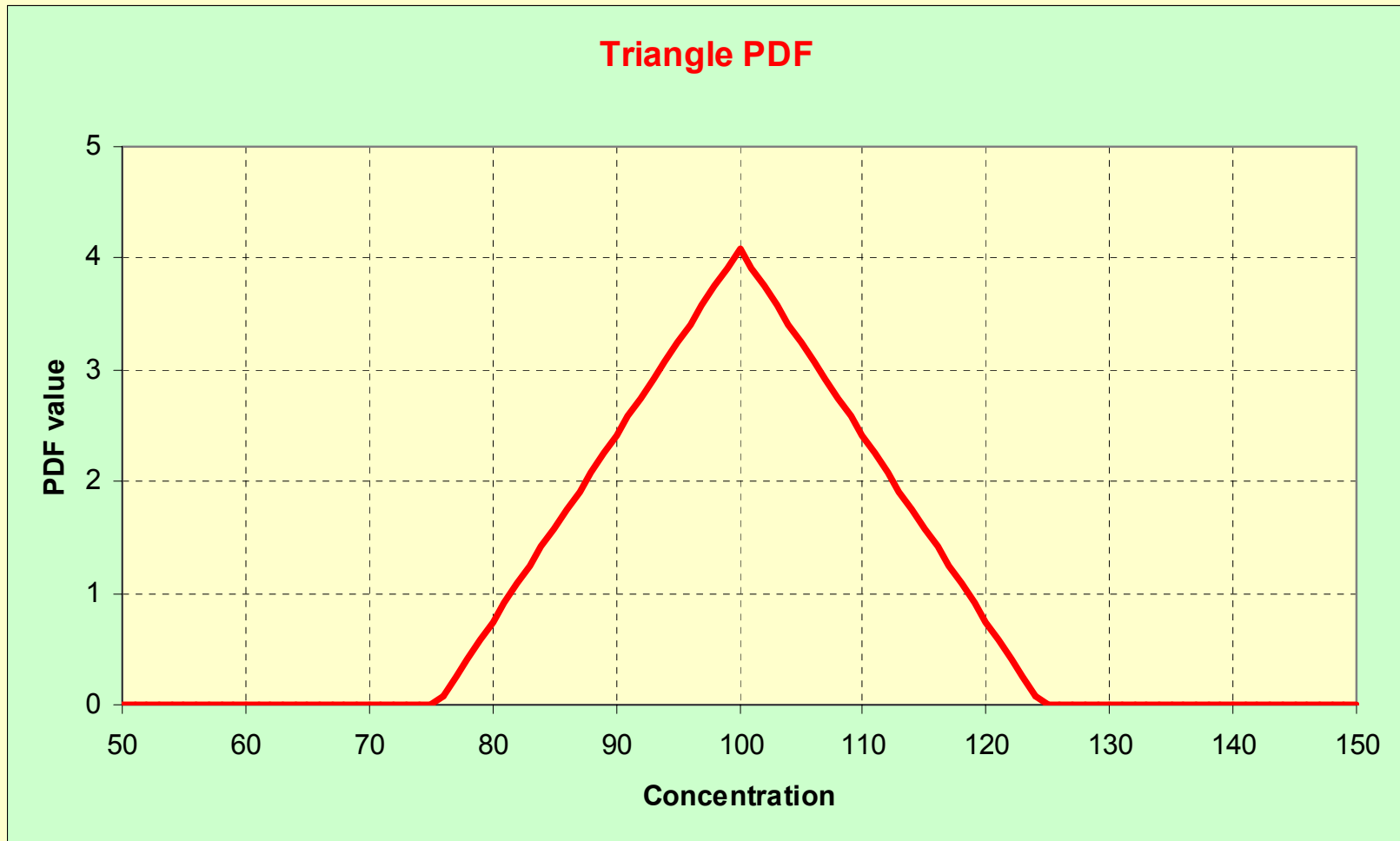
Air4EU DA tool: Philosophy

- The probability distributions (PDFs) describing the concentration uncertainties are denoted by p_L , p_U , p_R and p_O respectively and can be selected from the following five different types:
 - Uniform
 - Triangle
 - Gaussian (Normal)
 - Cauchy (Student's t or Lorentz)
 - Lognormal

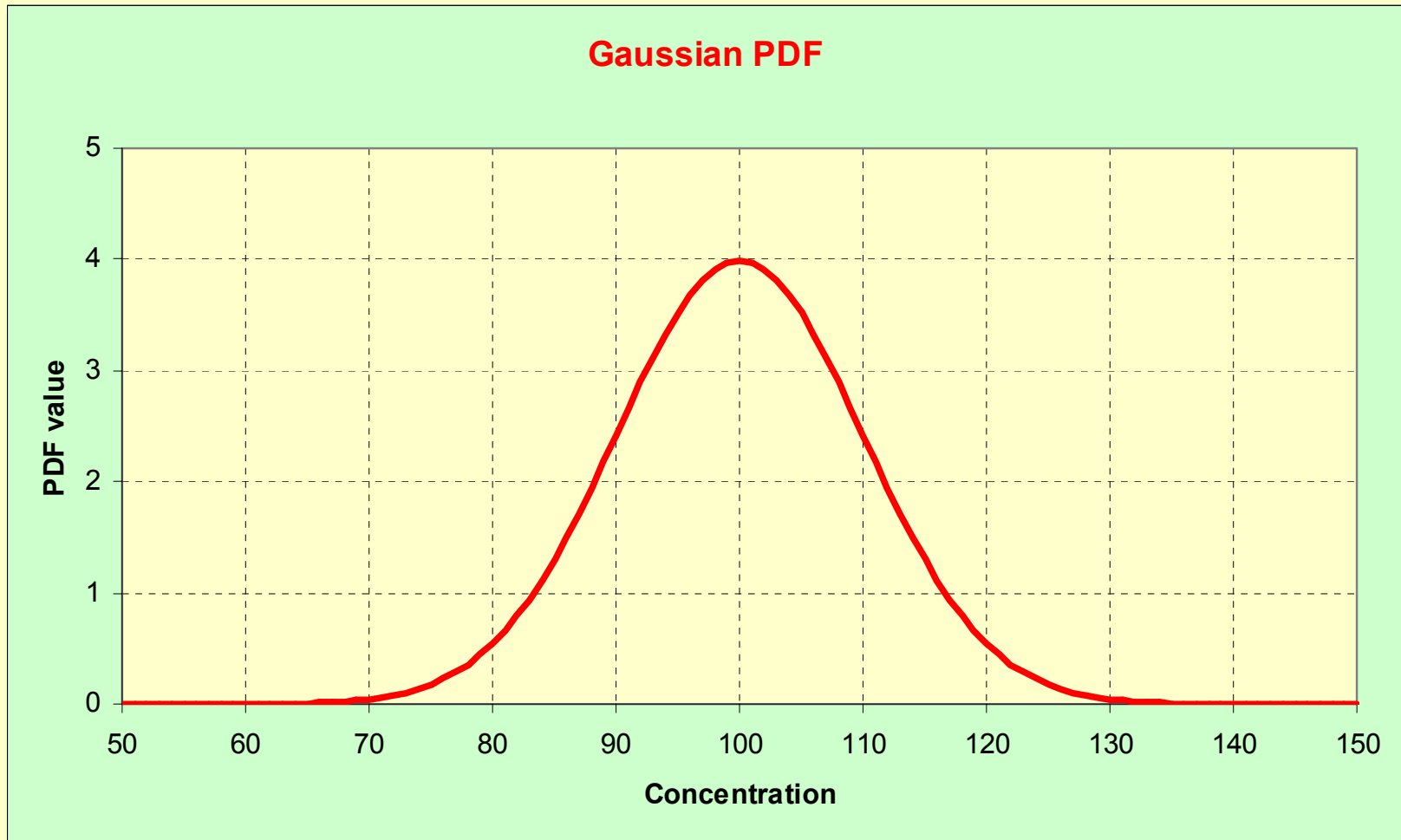
Air4EU DA tool: Uniform PDF



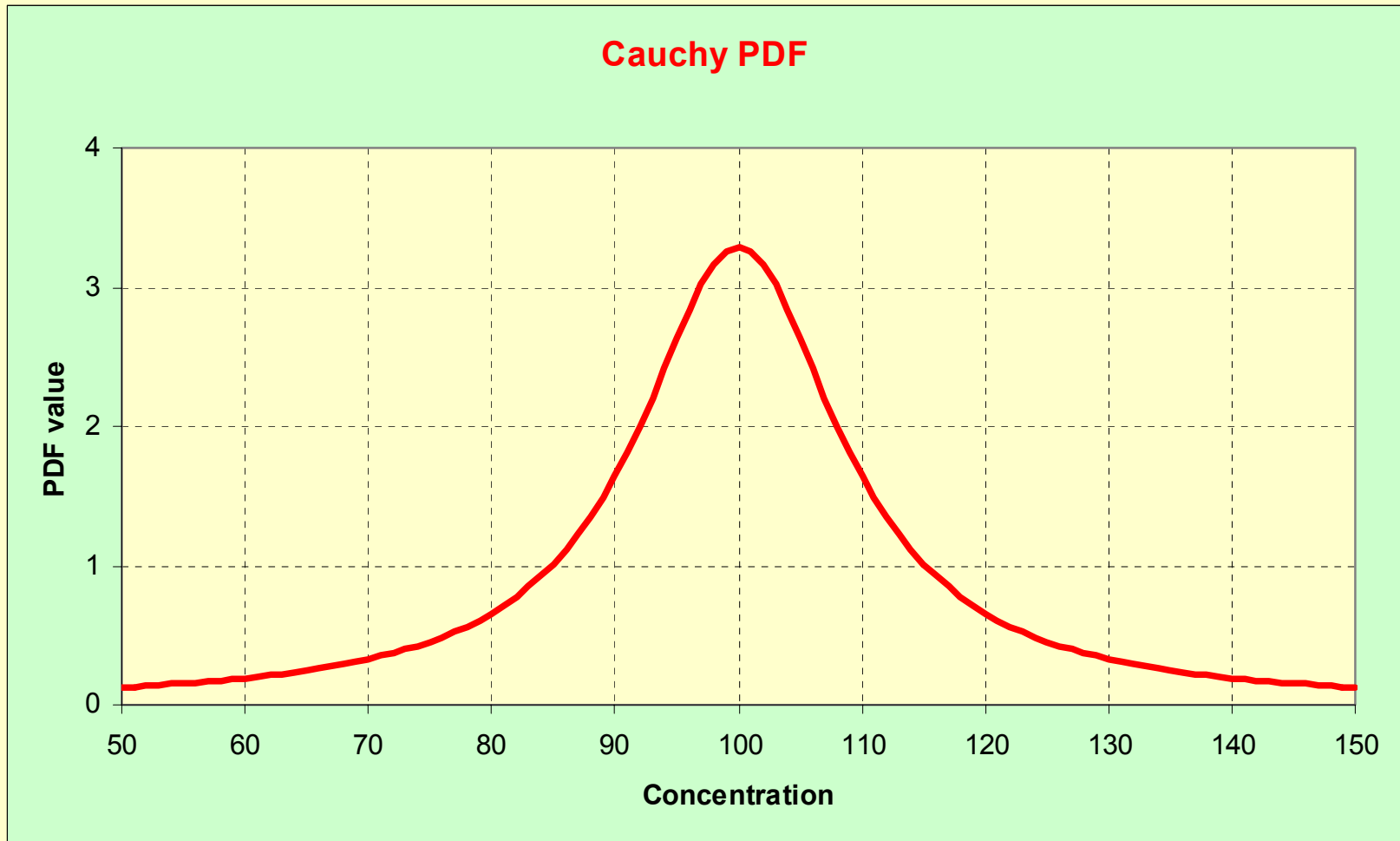
Air4EU DA tool: Triangle PDF



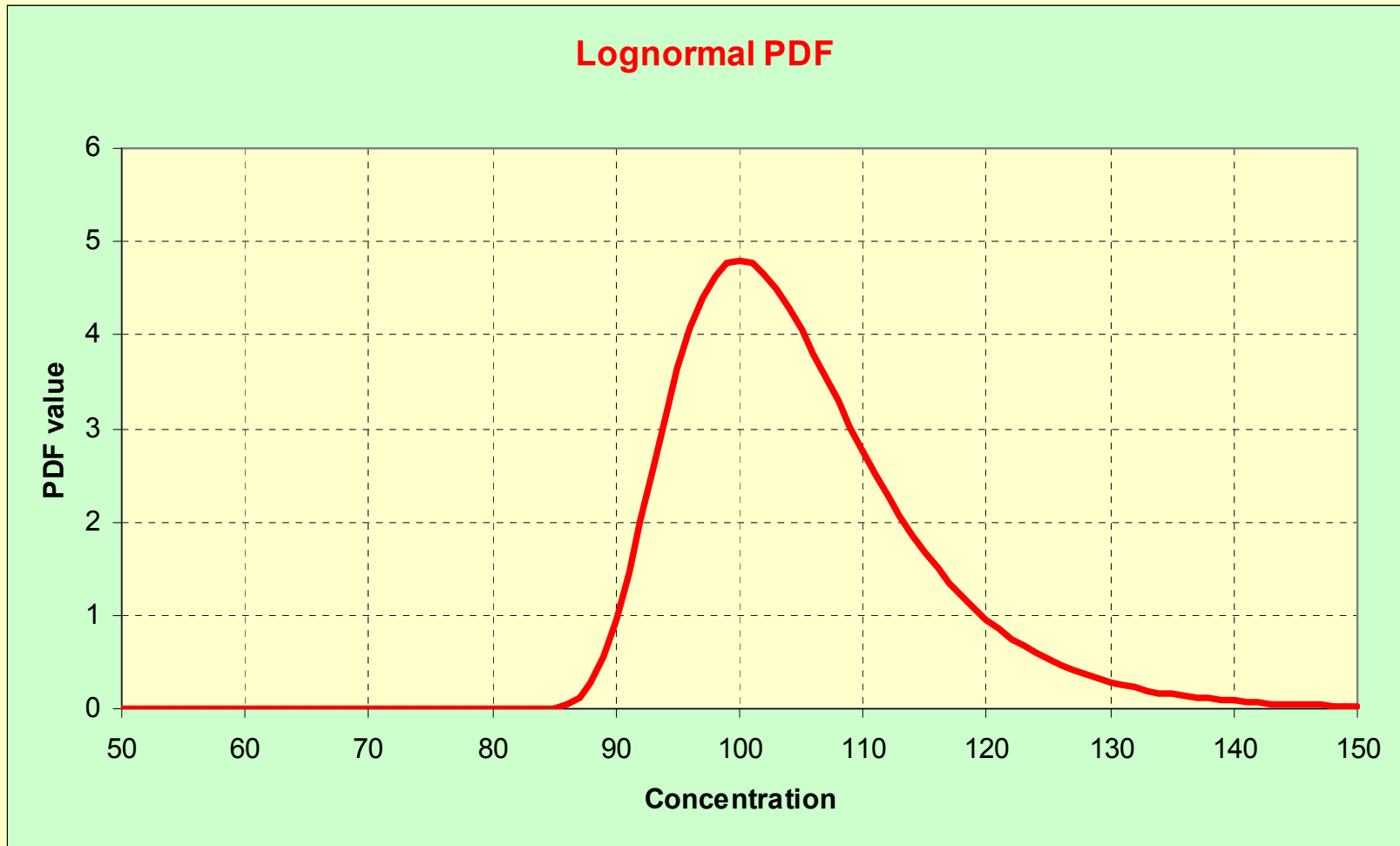
Air4EU DA tool: Gaussian PDF



Air4EU DA tool: Cauchy PDF



Air4EU DA tool: Lognormal PDF



Air4EU DA tool

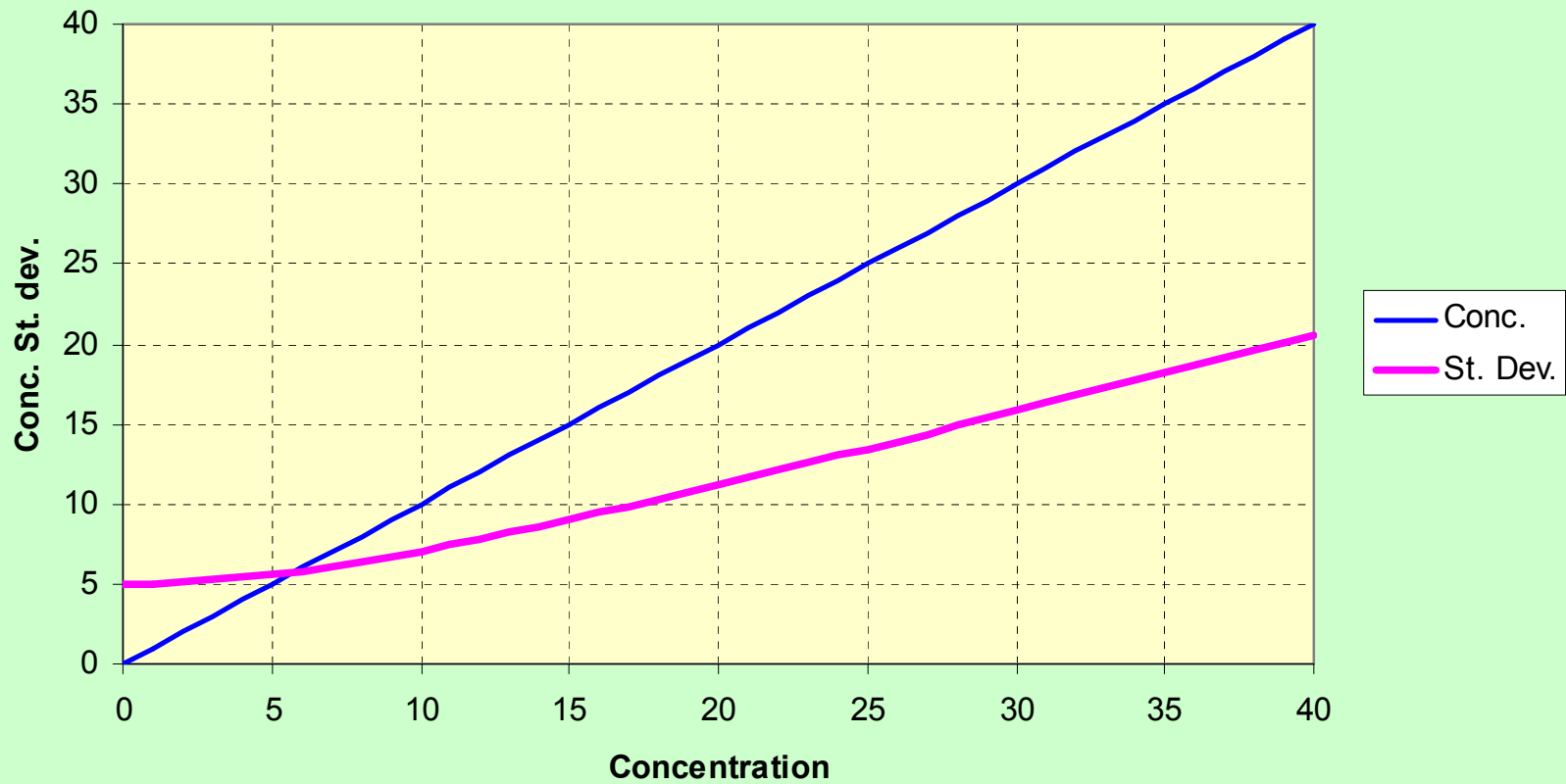
- The **standard deviation** σ_C for each PDF is calculated by the following formula:

$$\sigma_C = \sqrt{\sigma_A^2 + \sigma_R^2 \cdot C^2}$$

where σ_A and σ_R represents the **absolute and relative standard deviations** associated with the given concentration value C

Air4EU DA tool

Concentration and calculated standard deviation curve



Air4EU DA tool

- The Total Model concentration PDF p_M is calculated based on the PDFs p_L , p_U and p_R using the usual formula for **statistical convolution**:

$$p_M(c) = \int_{c_L} \int_{c_U} p_L(c_L) \cdot p_U(c_U) \cdot p_R(c - c_L - c_U) dc_L dc_U$$

We assume that the uncertainties associated with the scale contributions can be viewed as independent stochastic variables

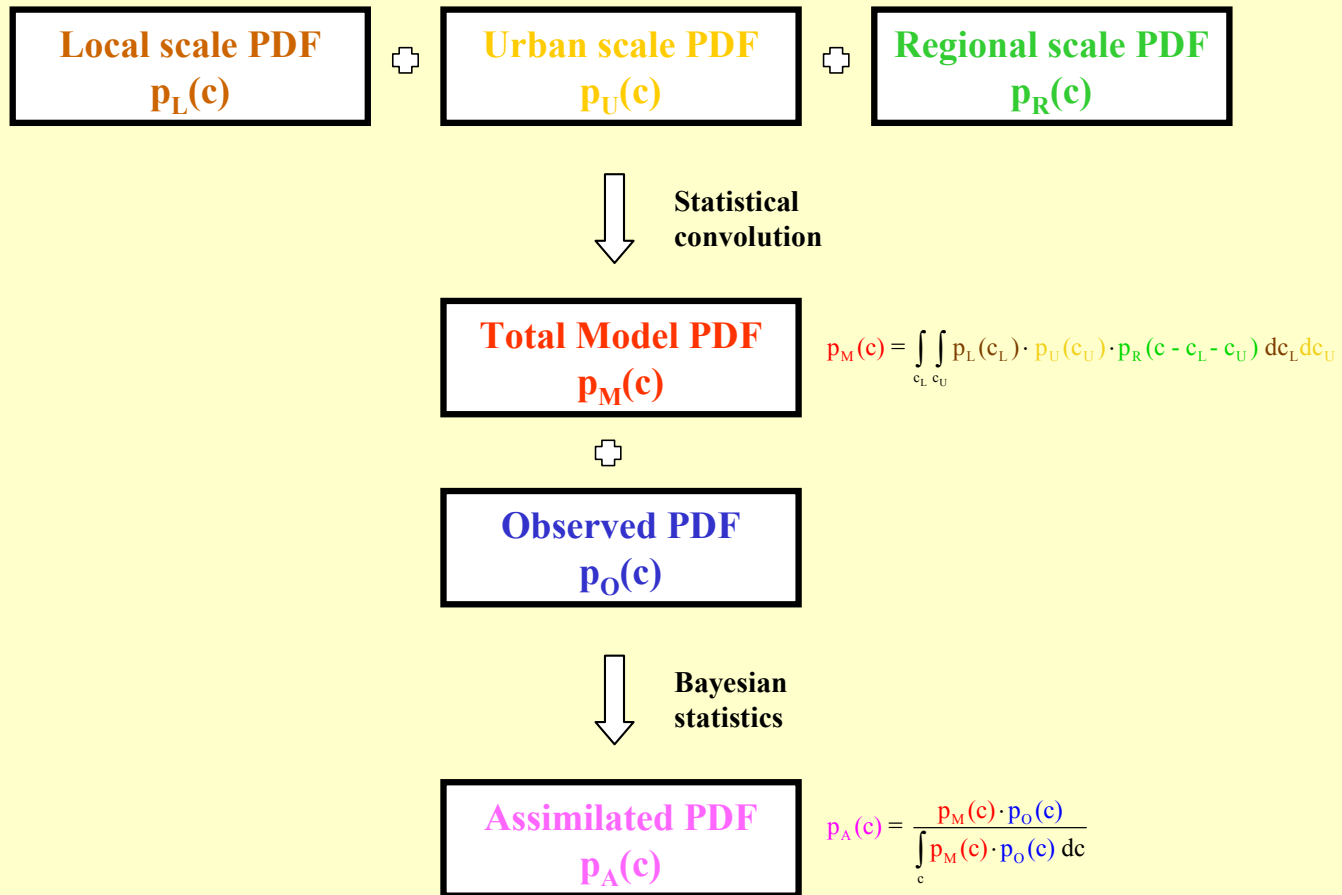
Air4EU DA tool

- The **data assimilation step** is performed by calculating a new assimilated concentration PDF p_A based on the Total Model PDF p_M and Observed concentration PDF p_O using Bayesian statistics:

$$p_A(c) = \frac{p_M(c) \cdot p_O(c)}{\int_c p_M(c) \cdot p_O(c) dc}$$

The posterior PDF p_A describes the probabilities of what the true underlying concentration C_T might be in the light of the assumptions and data provided

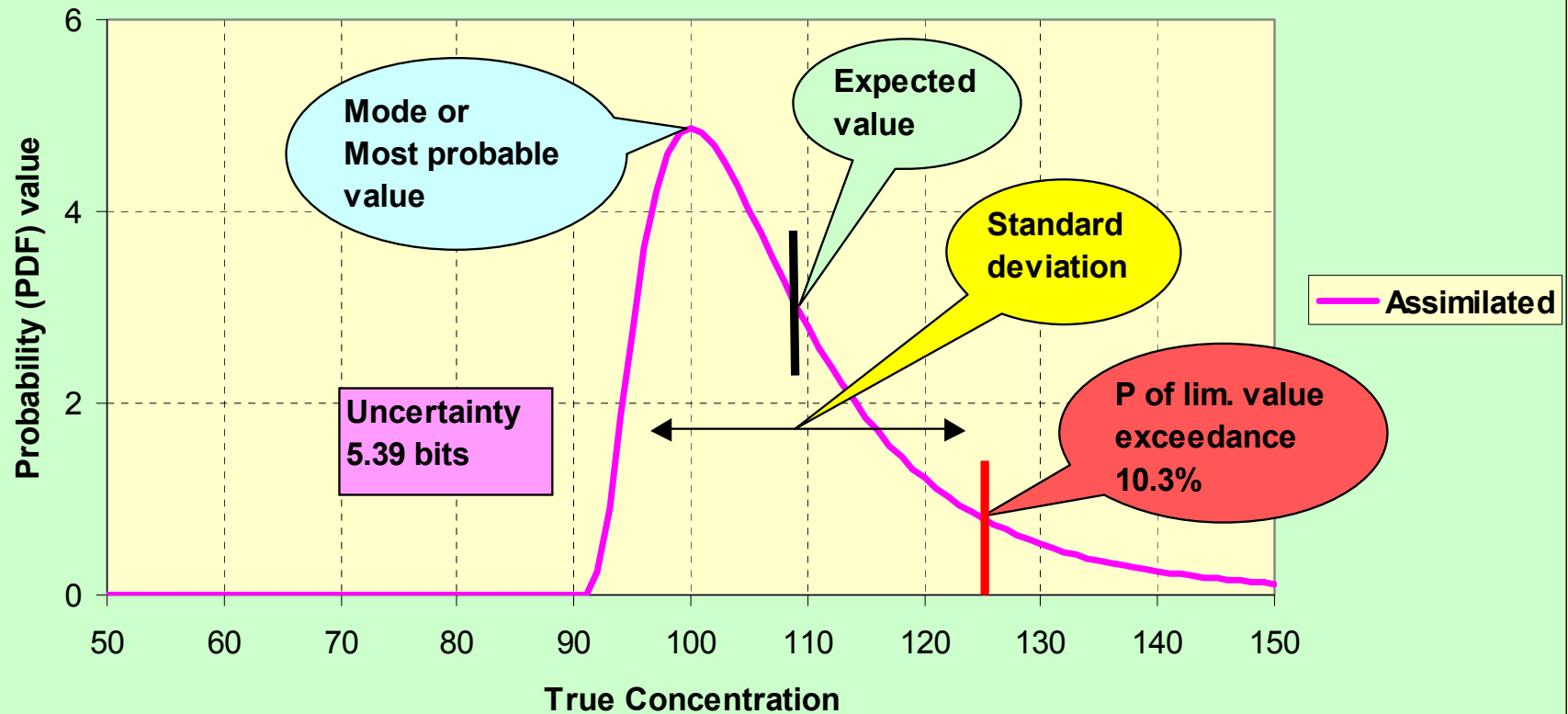
Overview of the Air4EU assimilation tool



Statistical inference about the true concentration

Air4EU DA tool: Output

Air4EU Simple Statistical System of Assimilation of Models and Monitoring data (SAM) Description of PDF output parameters



Air4EU DA tool: Entropy

- For Gaussian distributions the standard deviation is a good measure of uncertainty associated with a stochastic variable
- For other and more general distributions an uncertainty measure from the field of Information Theory known as the **Shannon entropy** $H(p)$ represents a better measure
- For a continuous probability distribution $p(c)$ the Shannon entropy $H(p)$ is defined by:

$$H(p) = - \int_c p(c) \cdot \log_2 p(c) \, dc$$

where the integral is taken over all values of c for which $p(c) > 0$

Air4EU DA tool: Entropy

- In the current version of the program the Shannon entropy is calculated based on a **discretization** of the distribution $p(c)$ with granularity 1 unit of concentration
- Thus in the program the entropy is calculated as:

$$H(p) = -\sum_{i=1}^n p_i \cdot \log_2(p_i)$$

where c_i represents the discrete concentration values with positive probability masses $p_i = p(c_i)$ for $i = 1, \dots, n$

- In the current version of the program $n = 10000$ and $c_i = \{0, 1, 2, 3, \dots, 9999\}$

Air4EU DA tool: Entropy

- **Assimilated vs. Total Model information gain** is defined as the reduction in uncertainty (entropy) obtained by data assimilation i.e., by replacing the Total Model concentration PDF p_M with the Assimilated concentration PDF p_A

Absolute Information Gain (bits):

$$\text{AIG}_{M \rightarrow A} = H(p_M) - H(p_A)$$

Relative Information Gain (%):

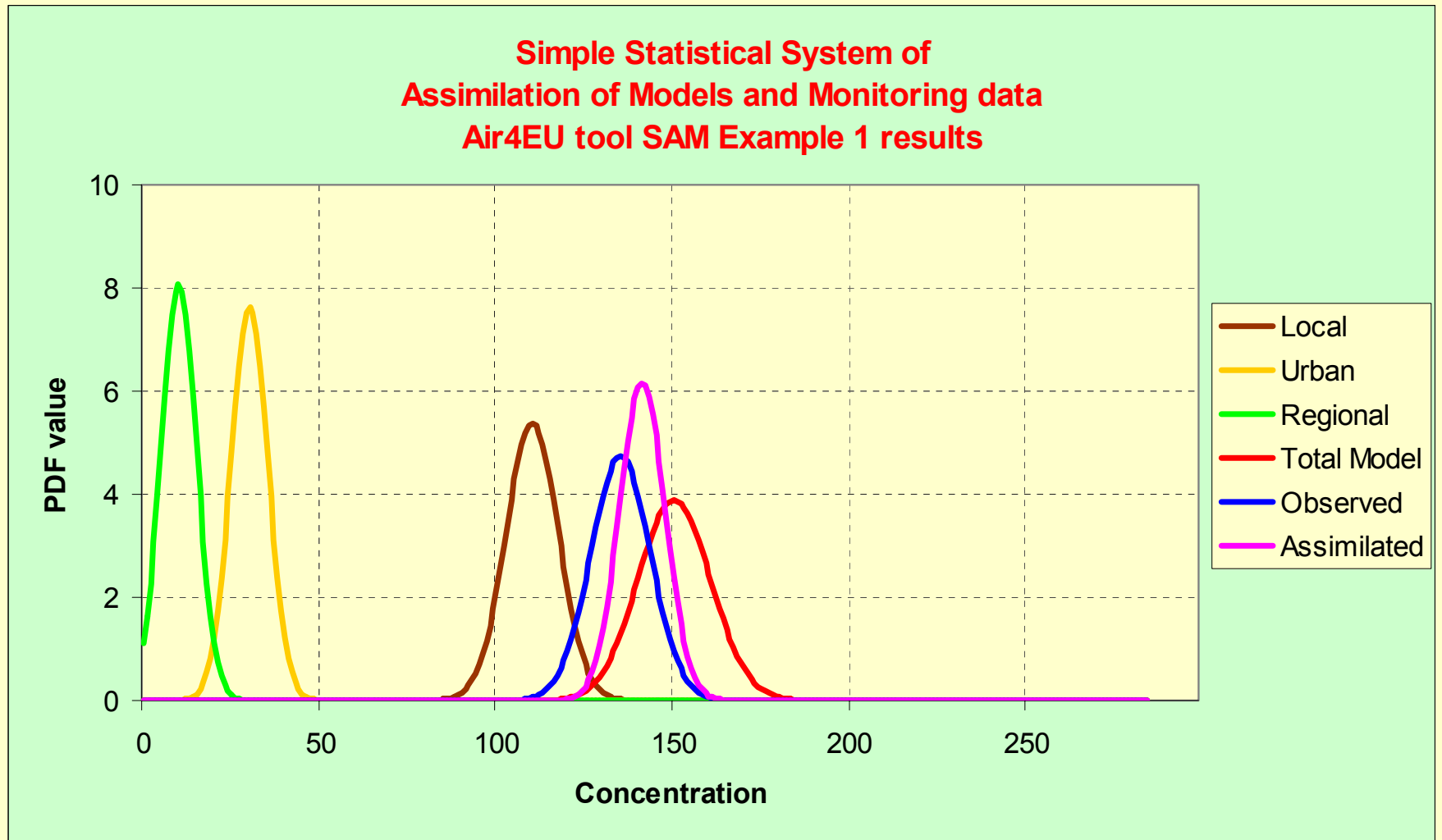
$$\text{RIG}_{M \rightarrow A} = 100 \cdot (H(p_M) - H(p_A)) / H(p_M)$$

Air4EU DA tool: Example 1 input

- Assume PDFs describing the concentration uncertainties associated with the different scales are defined as follows:
 - $p_L = \text{Gaussian}(110; 5., 0.05)$
 - $p_U = \text{Gaussian}(30; 5., 0.05)$
 - $p_R = \text{Gaussian}(10; 5., 0.05)$
 - $p_O = \text{Gaussian}(135; 5., 0.05)$

and a limit value is defined as $150 \mu\text{g}/\text{m}^3$

Air4EU DA tool: Example 1 results



Air4EU DA tool: Example 1 results

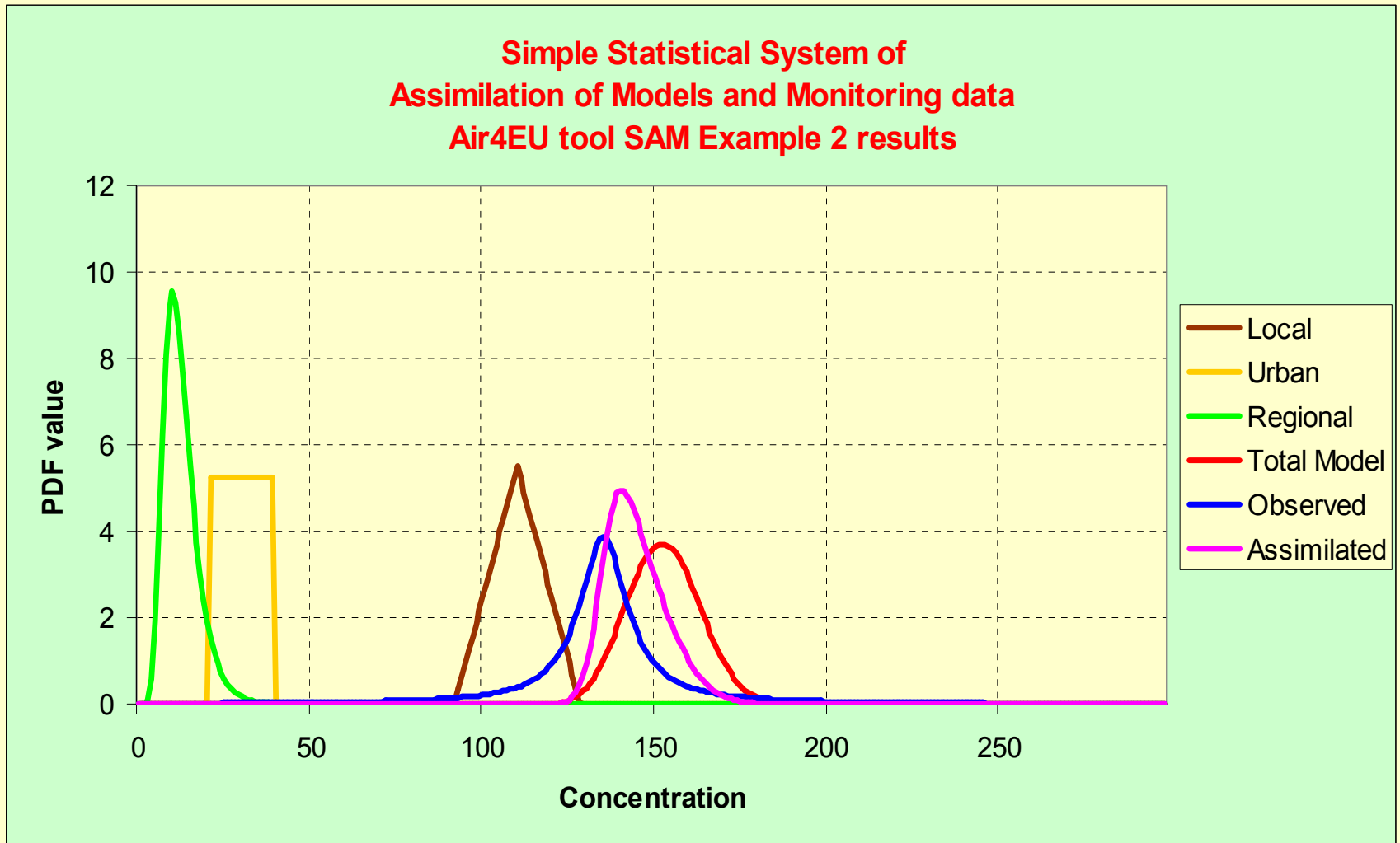
- The Total Model, Observed and Assimilated concentration PDFs are:
 - $p_M = \text{Gaussian}(150; 10.3)$
 - $p_O = \text{Gaussian}(135; 8.4)$
 - $p_A = \text{Gaussian}(141; 6.5)$
- The probability of exceeding the limit value (150) is calculated to be 7.4 %
- The relative information gain (RIG) is calculated to be 12.3%

Air4EU DA tool: Example 2 input

- Assume PDFs describing the concentration uncertainties associated with the different scales instead are defined as follows:
 - $p_L = \text{Triangle} (110; 5., 0.05)$
 - $p_U = \text{Uniform} (30; 5., 0.05)$
 - $p_R = \text{Lognormal} (10; 5., 0.05)$
 - $p_O = \text{Cauchy} (135; 5., 0.05)$

and a limit value again defined as $150 \mu\text{g}/\text{m}^3$

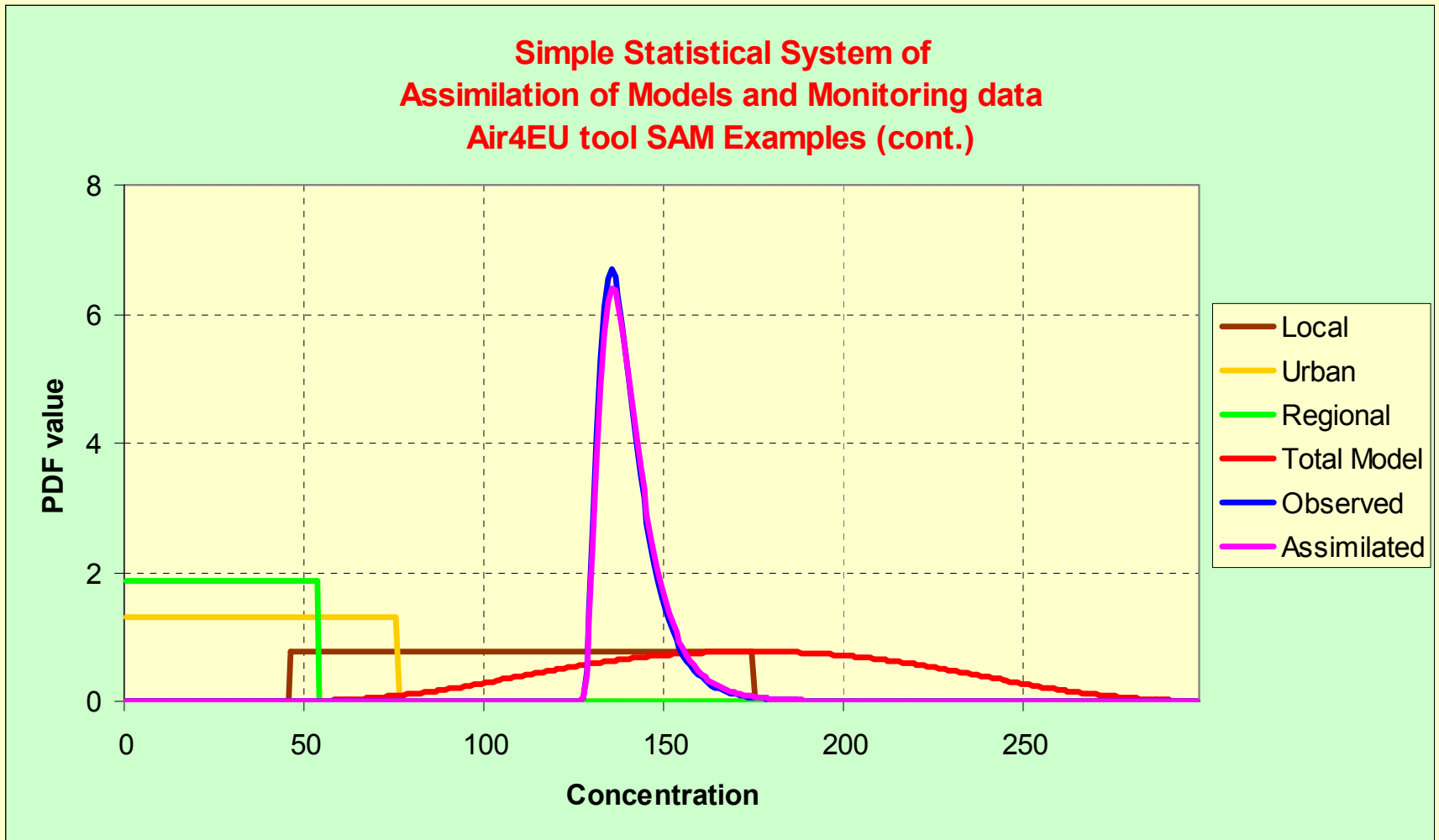
Air4EU DA tool: Example 2 results



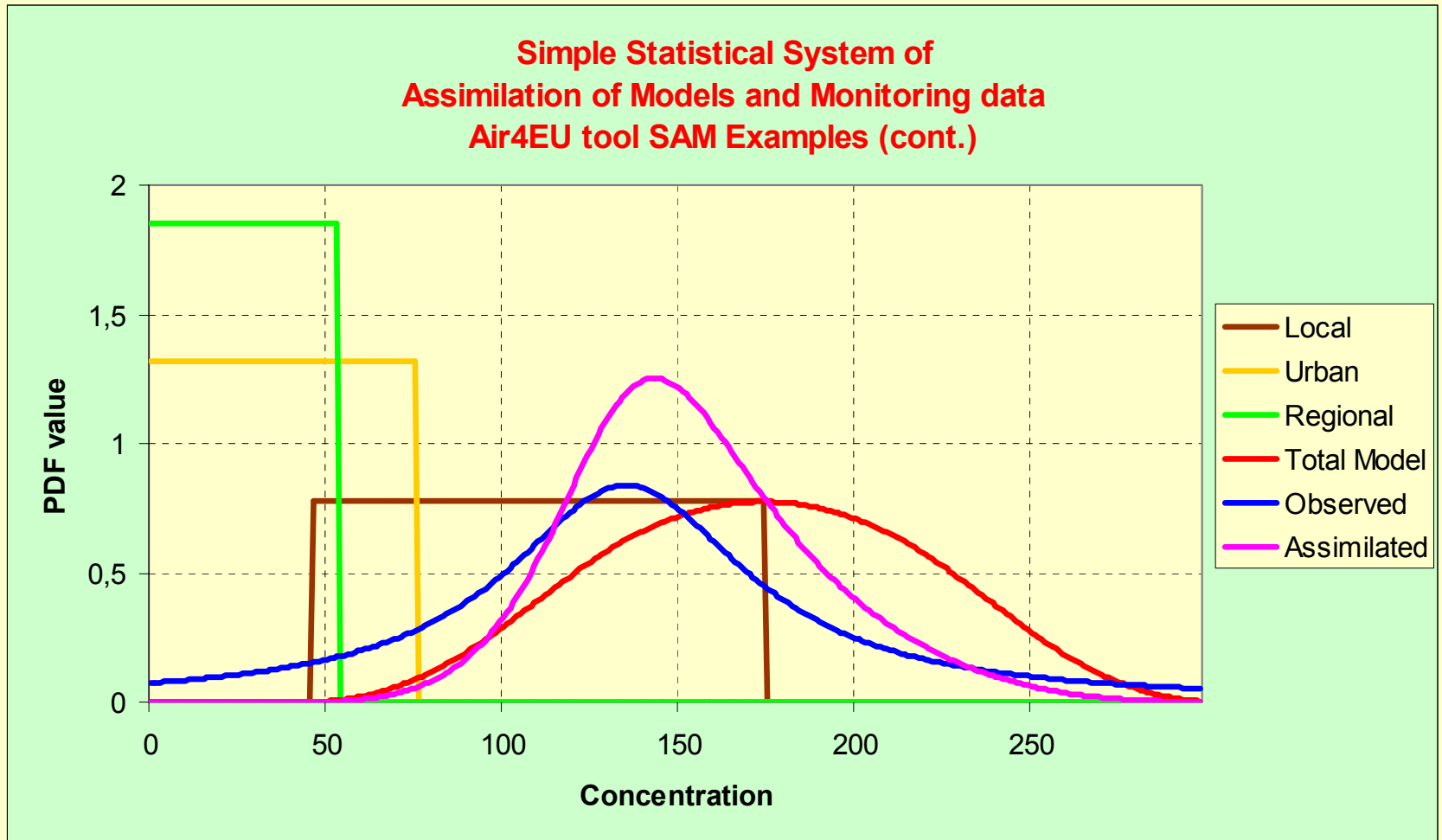
Air4EU DA tool: Example 2 results

- The Total Model, Observed and Assimilated concentration PDFs are:
 - $p_M = \text{Mixed}_1 (152.5; 10.5)$
 - $p_O = \text{Cauchy} (135; 8.4)$
 - $p_A = \text{Mixed}_2 (144.4; 8.8)$
- The probability of exceeding the limit value (150) is calculated to be 22.8 %
- The relative information gain (RIG) is calculated to be 5.7%

Air4EU DA tool: Examples cont.

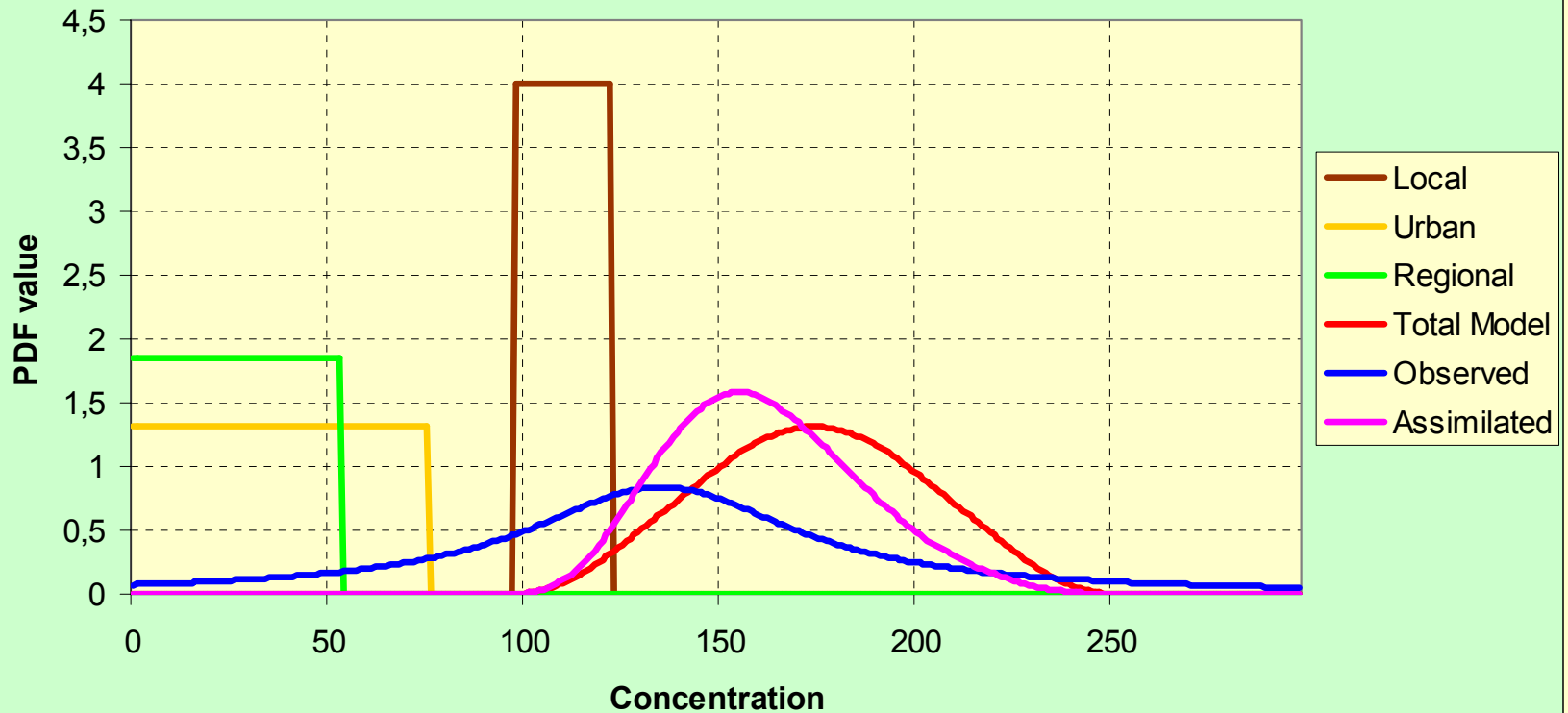


Air4EU DA tool: Examples cont.



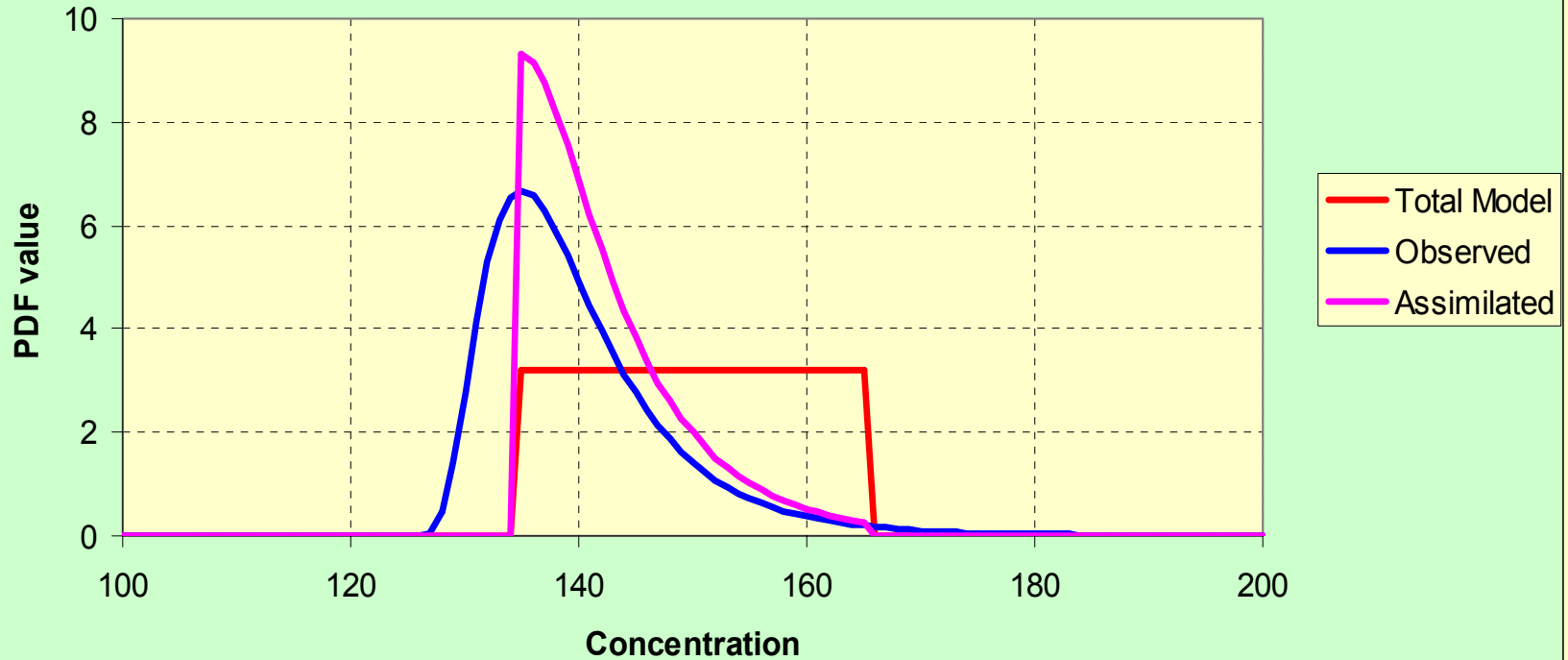
Air4EU DA tool: Examples cont.

Simple Statistical System of
Assimilation of Models and Monitoring data
Air4EU tool SAM Examples (cont.)



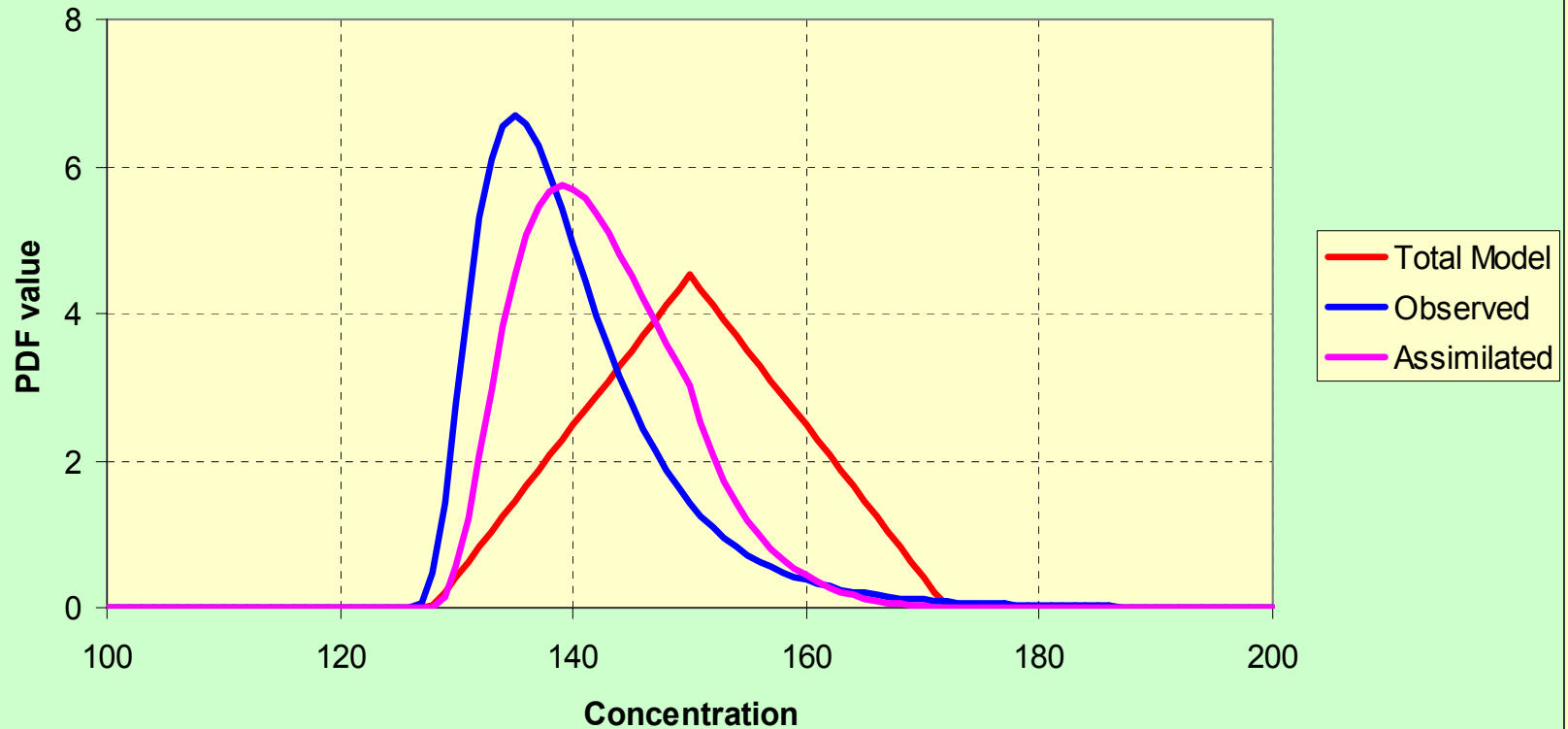
Air4EU DA tool: Examples cont.

**Simple Statistical System of
Assimilation of Models and Monitoring data
Air4EU tool SAM Examples (cont.)**

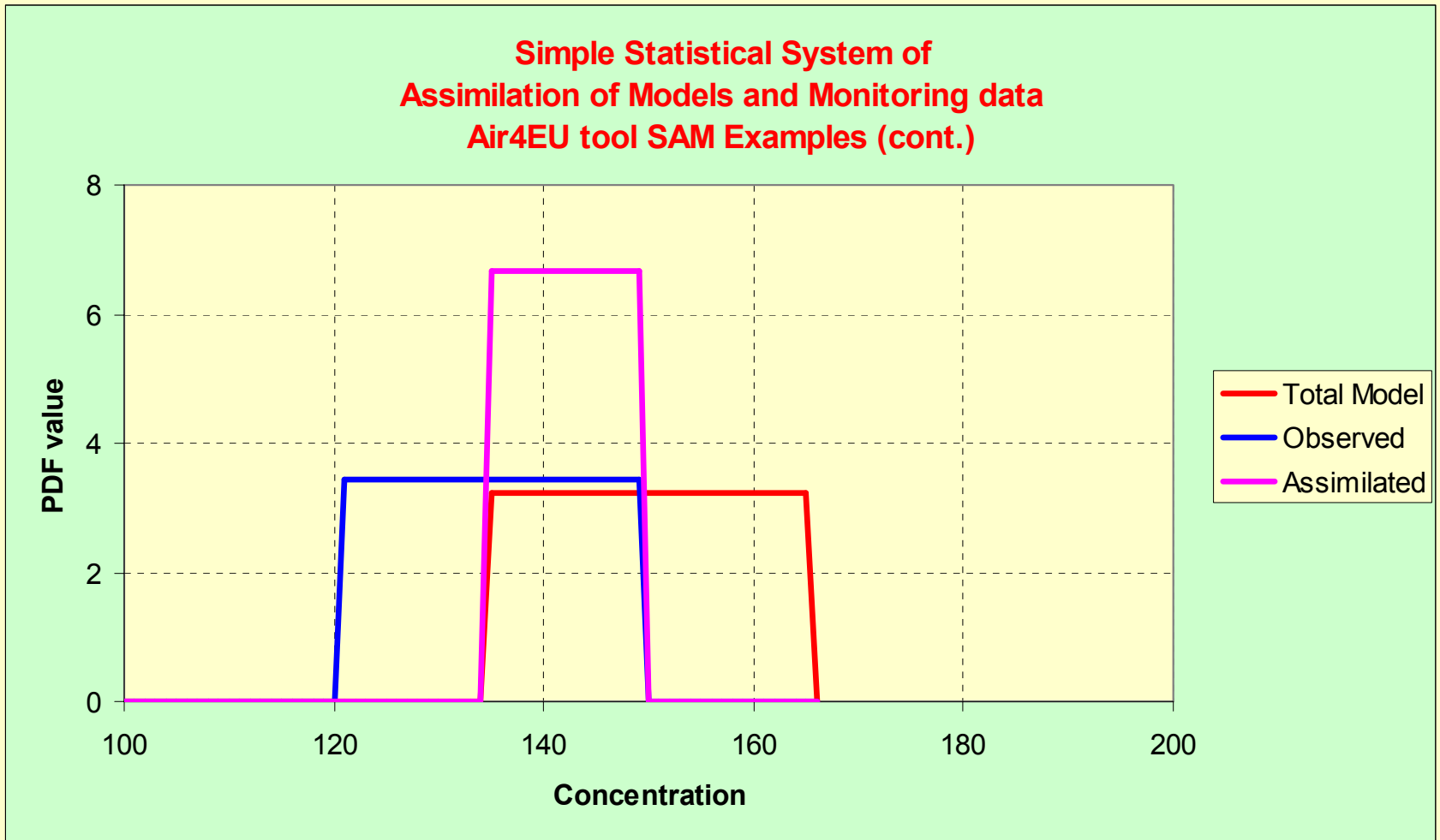


Air4EU DA tool: Examples cont.

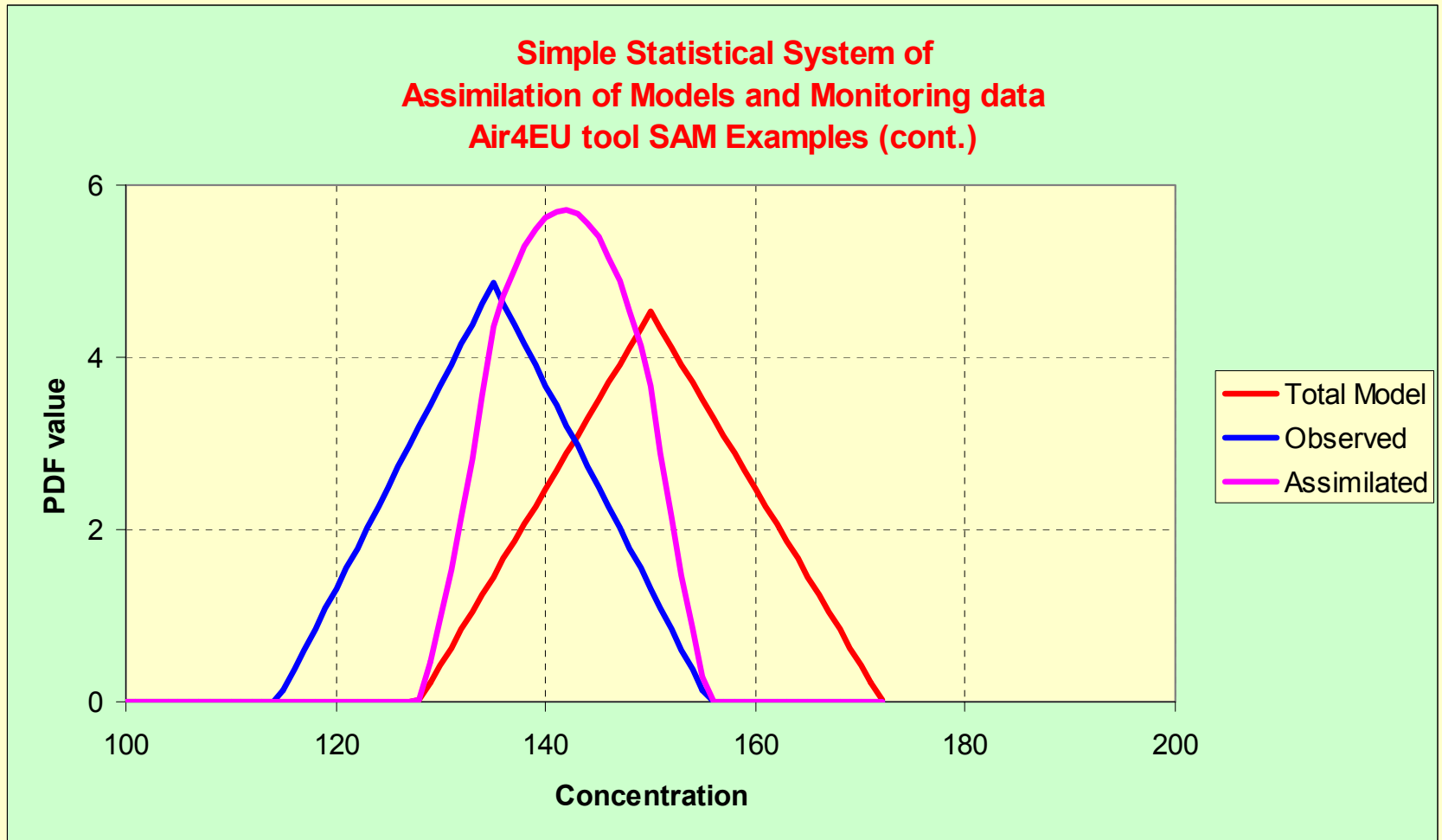
**Simple Statistical System of
Assimilation of Models and Monitoring data
Air4EU tool SAM Examples (cont.)**



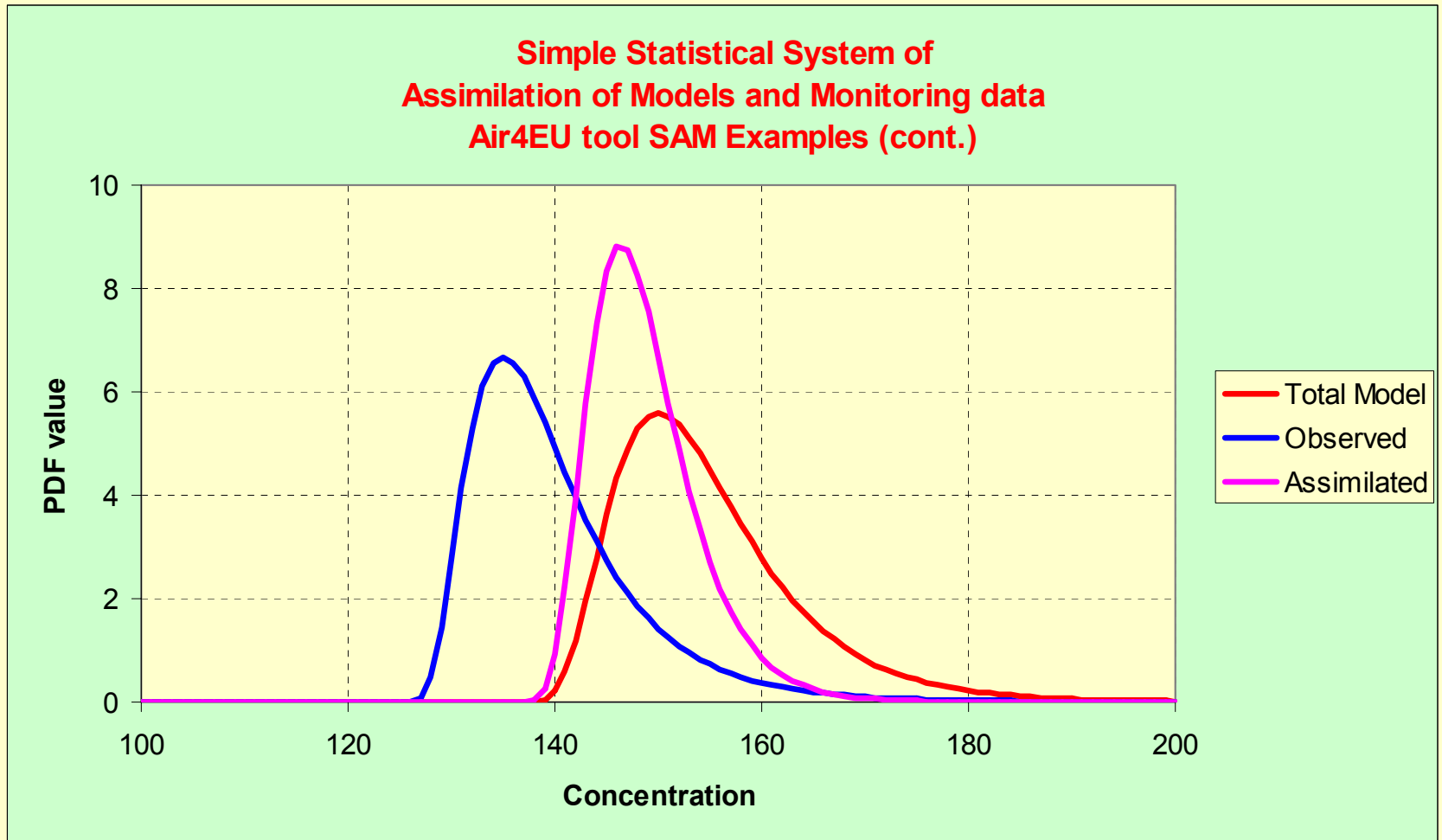
Air4EU DA tool: Examples cont.



Air4EU DA tool: Examples cont.

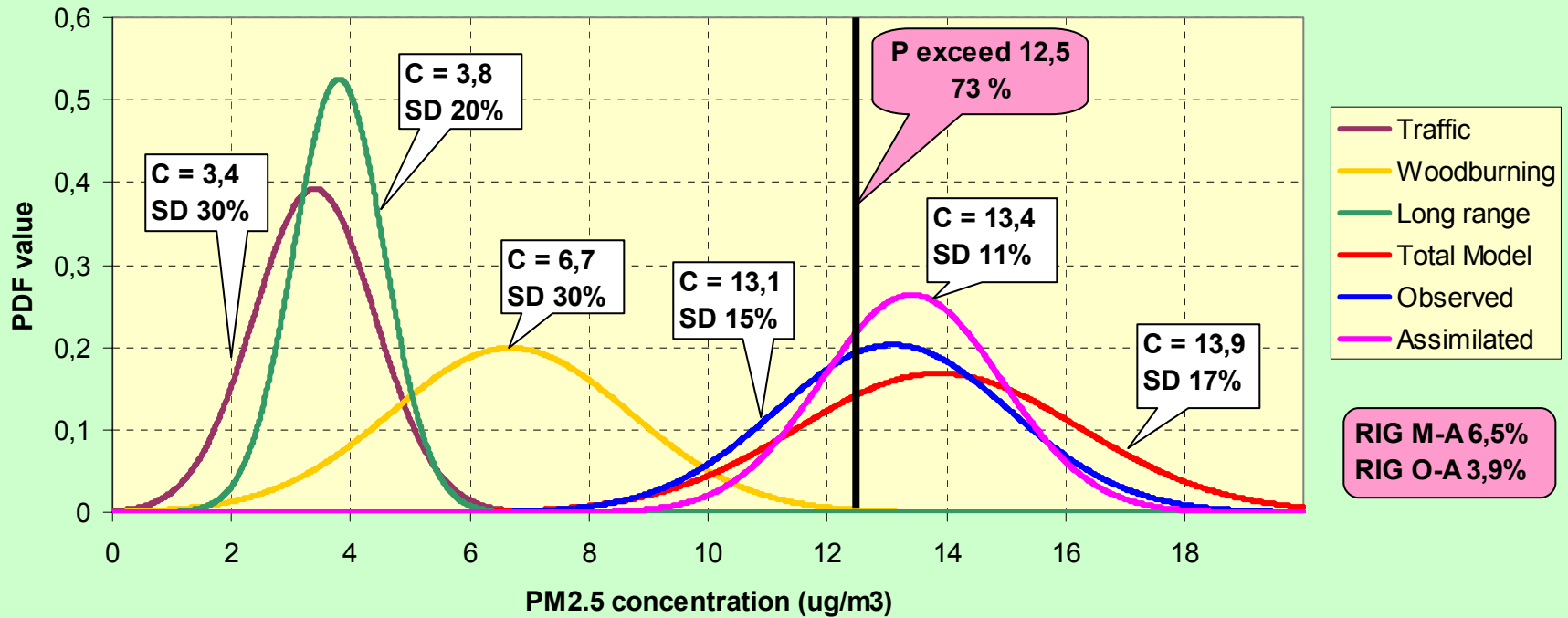


Air4EU DA tool: Examples cont.



Air4EU DA tool: Example from Oslo

PM2.5 concentration at station RV4 Aker hospital Oslo, Norway
Average over period 1.11.2003 - 30.4.2004



Air4EU DA tool: Future

- Can be made as a web application with graphics integrated
- Extension from a single point to an arbitrary set of spatial points (grids, irregular receptor points) and time periods
- Extension to sum more than 3 sets of model contributions and to an arbitrary number of observations
- Use of FFT to solve for the convolution!
- Keeping the possibility of operating with different marginal PDFs for each scale and spatial point
- Can use statistical copula functions to describe spatial dependencies which may lead to generalisations of present DA methods such as 3D-Var, EnKF etc.