A multi-fidelity framework is presented to accurately predict the turbulent Schmidt number, $S_{C_t}$, with applications to atmospheric dispersion modelling. According to the literature and experimental evidence, different physical correlations can be traced for $S_{C_t}$, relating this quantity to various turbulent parameters. The objective is to derive a reliable formulation for $S_{C_t}$ that can be used in various test cases and combined with several turbulence models in the context of Reynolds-averaged Navier-Stokes (RANS) simulations. To achieve that, high-fidelity data are obtained with a delayed Detached Eddy Simulation (dDES) and used in a correlation study to analyze the inter-dependencies of $S_{C_t}$ with important turbulent variables. A first data-driven model for $S_{C_t}$ is proposed by calibrating the data to the semi-empirical relation by Reynolds. A second model is presented using the results of a correlation study in combination with Principal Component Analysis (PCA). Both data-driven models were verified with the RANS simulation of the Cedval A1-5 case, and 2 additional dispersion cases: the Cedval B1-1 array of building, and the empty street canyon from the CODASC database. There can be concluded that the resulting $S_{C_t}$ formulation is able to significantly improve the accuracy of the concentration field compared to standard RANS approaches. Furthermore, the validity of the new formulation is demonstrated in combination with several turbulence models.

Keywords
Urban Environment; Atmospheric Boundary Layer; Turbulent Schmidt Number; Multi-fidelity framework; Data-driven; Principal Component Analysis

*Corresponding author’s email: riccardo.longo@ulb.ac.be ; alessandro.parente@ulb.be
## Nomenclature

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a, b, c, d, e, f$</td>
<td>coefficients in $Sc_t$ formulations</td>
</tr>
<tr>
<td>$A, B$</td>
<td>matrix of eigenvectors</td>
</tr>
<tr>
<td>$\bar{C}$</td>
<td>mean substance concentration, $ppm$</td>
</tr>
<tr>
<td>$C_0, C$</td>
<td>positive diffusion parameters</td>
</tr>
<tr>
<td>$C_{k1}, C_{k2}$</td>
<td>constant in the $k$ inlet profile</td>
</tr>
<tr>
<td>$C_1, C_2, m, n$</td>
<td>constants in $Sc_2$ formulation by Reynolds</td>
</tr>
<tr>
<td>$C_k$</td>
<td>substance concentration, $ppm$</td>
</tr>
<tr>
<td>$C_m$</td>
<td>measured tracer concentration, $ppm$</td>
</tr>
<tr>
<td>$C_s$</td>
<td>source tracer concentration, $ppm$</td>
</tr>
<tr>
<td>$C_n, C_{c1}, C_{c2}, \sigma$</td>
<td>constants in the $k - \epsilon$ model</td>
</tr>
<tr>
<td>$D, H, W, \lambda$</td>
<td>building’s dimensions, $m$</td>
</tr>
<tr>
<td>$D_m$</td>
<td>molecular diffusivity, $m^2 s^{-1}$</td>
</tr>
<tr>
<td>$D_t$</td>
<td>turbulent diffusivity, $m^2 s^{-1}$</td>
</tr>
<tr>
<td>$F_S$</td>
<td>safety factor</td>
</tr>
<tr>
<td>$k$</td>
<td>turbulent kinetic energy, $m^2 s^{-2}$</td>
</tr>
<tr>
<td>$K$</td>
<td>dimensionless concentration</td>
</tr>
<tr>
<td>$L$</td>
<td>vector of eigenvalues</td>
</tr>
<tr>
<td>$n$</td>
<td>number of observations</td>
</tr>
<tr>
<td>$Pe$</td>
<td>Péclet number</td>
</tr>
<tr>
<td>$Pr$</td>
<td>Prandtl number</td>
</tr>
<tr>
<td>$Q$</td>
<td>number of conserved variables</td>
</tr>
<tr>
<td>$Q_l$</td>
<td>emission rate of line source, $m^2 s^{-1}$</td>
</tr>
<tr>
<td>$Q_s$</td>
<td>total source flow rate, $m^3 s^{-1}$</td>
</tr>
<tr>
<td>$r_h$</td>
<td>coarsening ratio</td>
</tr>
<tr>
<td>$Re$</td>
<td>Reynolds number</td>
</tr>
<tr>
<td>$Re_\lambda$</td>
<td>Reynolds at the Taylor micro-scale</td>
</tr>
<tr>
<td>$Re_t$</td>
<td>turbulent Reynolds number</td>
</tr>
<tr>
<td>$S$</td>
<td>strain-rate invariant</td>
</tr>
</tbody>
</table>
$Sc$  molecular Schmidt number

$Sc_t$  turbulent Schmidt number

$u_x, u_y, u_z$  velocity components, $m \ s^{-1}$

$u_*$  ABL friction velocity, $m \ s^{-1}$

$x, y, z$  stream-wise, width and height coordinates, $m$

$x_i$  measured tracer molar fraction

$X$  matrix of samples

$\tilde{y}^+, y^+$  non-dimensional wall distances

$z_0$  aerodynamic roughness length, $m$

$\delta_u, \delta_k, \delta_h$  local deviation of turbulent properties

$\epsilon$  turbulent dissipation rate, $m^2 s^{-3}$

$\kappa$  von Karman constant

$\mu_t$  dynamic turbulent viscosity, $kg \ m^{-1} s^{-1}$

$\nu$  kinematic molecular viscosity, $m^2 s^{-1}$

$\nu_t$  kinematic turbulent viscosity, $m^2 s^{-1}$

$\omega$  specific rate of turbulence dissipation, $s^{-1}$

$\Omega$  vorticity invariant

$\rho$  density, $kg \ m^{-3}$
1. Introduction

The study of atmospheric dispersion is of major relevance to assess the air quality in urban environments. The major sources to urban air pollution are traffic (25%), combustion and agriculture (22%), domestic fuel burning (20%), natural dust (18%) and industrial activities (15%) [1]. Atmospheric dispersion models can be employed to determine the consequences of accidental chemical releases [2, 3, 4, 5, 6, 7]. Computational fluid dynamics (CFD) has been acknowledged as a powerful tool to predict urban flows [8] and dispersion patterns [9, 10, 11, 12, 13]. CFD dispersion modelling requires the definition of $Sc_t$, expressing the ratio of turbulent viscosity to mass diffusivity. However, the literature does not report clear guidelines on the definition of this parameter [14, 15, 16, 17]. Nevertheless, the importance of $Sc_t$ was largely demonstrated [18, 19, 20, 21]. For Atmospheric Boundary Layer (ABL) flows, typical values range between 0.2 and 1.3 [22]. Reynolds [23] proposed an empirical formulation for the definition of $Sc_t$ in function of the molecular $Sc$ and the eddy viscosity ratio. The local variability of $Sc_t$ is supported by experimental evidence [24, 25], suggesting that $Sc_t$ should be prescribed as a dynamic variable [26, 27, 28]. In this regard, Gorlé et al. [29] developed a local $Sc_t$. The latter depends on $C_\mu$ and on the Reynolds number at the Taylor micro-scale. Subsequently, Longo et al. [15] proposed a $Sc_t$ formulation depending on $C_\mu$ whose main dispersion parameters were optimized through uncertainty quantification. Di Bernardino et al. [25] found a reasonable agreement between their $Sc_t$ measurements and the outcomes of the $Sc_t$ formulation by Longo et al. [15]. The experimental observations [25] and the latest variable $Sc_t$ formulations [29, 15] open the way to the present research.

A multi-fidelity framework combines computationally expensive high-fidelity simulations with simplified lower-order approximations to achieve affordable predictions at a desired user-defined accuracy [30, 31]. In the present framework, high-fidelity data are obtained using delayed Detached Eddy Simulations (dDES) and processed to develop a reliable model for $Sc_t$. The data-driven formulation for $Sc_t$ is used afterwards in Reynolds-Averaged Navier-Stokes (RANS) simulations, characterized by a lower fidelity. Two strategies are investigated to develop a data-driven model for $Sc_t$. The first method calibrates the high-fidelity data to a prescribed model form, following the semi-empirical formulation by Reynolds [23]. The second strategy relies on Principal Component Analysis (PCA) to optimally correlate $Sc_t$ to a significant number of model parameters. PCA is a processing technique that allows to extract the main features in large multivariate data structures. The method can be used to derive a smaller set of correlated variables containing most of the variance of
the system. This subset of components can be used to retrieve and interpret important flow features or build reduced-order models [32, 33, 34]. Coussement et al. [35] proposed a technique, the Manifold Generated PCA (MG-PCA) method, where a subset of the original state-space variables is retained. The correlations between all variables is retrieved by using the information contained in the PCA matrices. In the present investigation, MG-PCA is used to optimally correlate a subset of relevant turbulent quantities involved in the dispersion phenomenon with $S_{cl}$. It will be demonstrated how the MG-PCA based $S_{cl}$ formulation is able to improve the accuracy of dispersion prediction compared to standard approaches.

The novelty of this paper resides in the use of a comprehensive data-driven approach starting from high-fidelity data obtained with a dDES to derive a generalized $S_{cl}$ formulation. The latter can be successfully employed in conjunction with different RANS methodologies without limiting its local variability.

The paper is structured as follows: Section 2 details the turbulence and dispersion modelling of the ABL and presents the simulation test cases (Figure 1): (1) Cedval A1-5, a single building, (2) B1-1, an array of buildings [36], and (3) an empty street canyon from the CODASC database [37]. Section 3 explains the basics of principal component analysis. The results of the present data-driven approach in a multi-fidelity framework are presented in Section 4. The results section starts with the calibration of the high-fidelity data to the semi-empirical formula of Reynolds [23]. Next, an optimized formulation for $S_{cl}$ is derived after a correlation analysis and the use of MG-PCA. Finally, the resulting $S_{cl}$ formulations are validated on different cases in combination with various turbulence formulations. The novel approach demonstrates superior performance albeit its coupling to turbulence models that are not optimized for ABL flows. Conclusions are drawn in Section 5.
Figure 1: Cedval A1-5 single building (a-b), B1-1 array of buildings (c-d) and empty street canyon (e-f) test cases. The Cedval single building, the training case for this study, has been marked in red. Subfigures b-d-f display a zoomed view of the source buildings and the pollutant emitting sources (red squares/rectangles). Four sources are located on the leeward side of the emitting buildings for the test cases A1-5/B1-1 (b-d) while four sources are located on the ground, in-between the buildings, for the empty street canyon (f). For all the test cases under study, the pollutant is released simultaneously.
2. Turbulence and dispersion modelling

The CFD investigation of the atmospheric boundary layer is typically carried out by either solving the RANS equations either performing Large-Eddy Simulations (LES). Although LES simulations can lead to better results, they are at least one order of magnitude more expensive than RANS [38, 39]. A compromise between the two approaches is represented by Detached-Eddy Simulations (DES), where unsteady RANS models are employed in the boundary layer, switching to LES treatment in the separated regions. In the present work, DES is employed to obtain high fidelity data, to retrieve a novel formulation of the $S_{C_{1}}$. Subsequently, the resulting $S_{C_{1}}$ is employed and assessed coupled to a modified $k – \epsilon$ closure, designed for ABL flows [40, 41, 39, 42]. It employs a set of boundary conditions and turbulence variables, as expressed in Table 1. The modelling approach is completed by a wall formulation based on the local aerodynamic roughness [39, 43]. Whenever an obstacle is involved, a Building Influence Area (BIA), i.e. an area where non-linear eddy-viscosity models [44, 45, 46, 47] are applied, is used [40, 39, 43]. The BIA is able to automatically envelope the obstacles immersed in the flow-field.

As for the DES settings, the delayed Detached Eddy Simulation (dDES), proposed by Spalart et al. [48], was employed. With respect to the standard DES, dDES uses a modified length scale to remedy the problems of grid-induced separation (GIS) and modelled stress depletion (MSD) [49]. The dDES model chosen for this study is based on the realizable $k – \epsilon$ model. This permitted to use the same boundary conditions imposed for the RANS simulation. In this model the dissipation term is modified as follows:

$$Y_k = \frac{p k^{\frac{5}{3}}}{l_{des}},$$

with:

$$l_{des} = \min(l_{rke}, l_{les}),$$

$$l_{rke} = \frac{k^{\frac{3}{2}}}{\epsilon},$$

Table 1: Set of inlet conditions and turbulence variables for the "comprehensive approach" [39].

<table>
<thead>
<tr>
<th>Inlet Conditions</th>
<th>Turbulence Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U = \frac{u_s}{\kappa} ln \left( \frac{z+z_0}{z_0} \right)$</td>
<td>$\mu_t = C_{\mu} \rho \frac{k^2}{\epsilon}$</td>
</tr>
<tr>
<td>$k(z) = C_{k1} ln (z + z_0) + C_{k2}$</td>
<td>$S_k(z) = \frac{\rho u_t^4}{(z + z_0)^3} \left( \frac{(C_{\epsilon2} - C_{\epsilon1}) \sqrt{C_{\mu}}}{\kappa^2} - \frac{1}{\sigma_\epsilon} \right)$</td>
</tr>
<tr>
<td>$\epsilon(z) = \frac{u_s^3}{\kappa (z + z_0)}$</td>
<td>$C_{\mu} = \frac{u^4}{k^2}$</td>
</tr>
</tbody>
</table>
\[ l_{\text{des}} = C_{\text{des}} \Delta, \]  

(4)

where \( C_{\text{des}} \) is a calibration constant used in the DES model with a value of 0.61. \( \Delta \) is the maximum local grid spacing: \( \Delta \equiv (\Delta x, \Delta y, \Delta z) \). In the event that \( l_{\text{des}} = l_{rke} \), the expression for the dissipation of the \( k \) formulation for the Realizable \( k-\epsilon \) model is restored:

\[ Y_k = \rho \epsilon. \]  

(5)

In the delayed DES approach, the DES length \( l_{\text{des}} \) can be finally redefined according to:

\[ l_{\text{des}} = l_{rke} - f_d \max(0, l_{rke} - C_{\text{des}} \Delta). \]  

(6)

As for the concentration field, the dispersion of a pollutant is calculated solving a scalar transport equation for the solute concentration, \( C \). Referring to the Standard Gradient Diffusion Hypothesis (SGDH), the concentration transport equation is reduced to the following conservation equation for RANS approaches:

\[ \frac{\partial \bar{C}}{\partial t} + \frac{\partial}{\partial X} \left( \bar{U} \bar{C} - (D_m + D_t) \frac{\partial \bar{C}}{\partial X} \right) = \bar{\dot{C}}, \]  

(7)

or for LES approaches:

\[ \frac{\partial \tilde{C}}{\partial t} + \frac{\partial}{\partial X} \left( \tilde{U} \tilde{C} - (D_m + D_t) \frac{\partial \tilde{C}}{\partial X} \right) = \tilde{\dot{C}}, \]  

(8)

where dashes and swung dashes express mean and filtered quantities, with \( D_m = \nu / Sc \) the molecular diffusion coefficient and \( D_t = \nu_t / Sc_t \) the turbulent diffusion coefficient.

Significant uncertainty is associated to the specification of \( Sc_t \). Reynolds [23] proposed a variable formulation based on the molecular Schmidt number and the eddy viscosity ratio:

\[ Sc_t = C_1 \exp \left[ -C_2 Sc \left( \frac{\nu_t}{\nu} \right)^m \right], \]  

(9)

with \( C_1, C_2, m \) and \( n \) empirically defined constants.

Gorlé et al. [51] demonstrated that better results are achieved when adopting a variable \( Sc_t \) formulation, based on the solution obtained for the flow-field. Previous work by Longo et al. [15] supported this per-
spective and proposed the following local definition for $S_{c_t}$:

$$S_{c_t} = \frac{2C_\mu}{C_0C^2},$$

with $C_0 = 2$ and $C = 0.35$, two model parameters optimized through uncertainty quantification (UQ).

Dispersion processes are complex phenomena usually studied through simplified configurations [52]: the isolated building, the street canyon and the array of buildings.

In the present study, three test cases provided with experimental data were considered. The Cedval A1-5 single building [36], the training case, on which the data-driven derivation of a novel $S_{c_t}$ formulation was carried out, employing a dDES. Subsequently, the $S_{c_t}$ model was verified in the RANS framework on the same Cedval A1-5 and on two supplementary test cases: the Cedval B1-1 cluster of building [36] and the empty street canyon from the CODASC database [37] (Figure 2).
Figure 2: Test cases under study, including the profile lines taken into consideration for concentration measurements: (a) Cedval A1-5 [39], marked as the training case with a red square. (b) Cedval B1-1. (c) CODASC empty street canyon.
A brief description of the cases under study follows:

**Cedval A1-5 single building** - The Cedval A1-5 is the training case, selected for the data-driven derivation, based on the outcomes of a dDES. It features a scaled 1 : 200 rectangular single building (Figure 2 (a)).

As shown in Figure 1 (b), the building’s dimensions are: $L = 0.1m$, $W = 0.15m$ and $H = 0.125m$. The pollutant is emitted simultaneously from 4 source elements located on the leeward facade of the buildings, with a fixed velocity of $U = 0.024 [m/s]$. The roughness length and friction velocity were specified equal to 0.0007m and 0.3777m/s, respectively. Dimensions and precise locations of the source terms are specified in the supplementary material. The origin of the coordinate system is at the center of the bottom face of the building. The $z$-axis points upwards while the $x$-axis points downstream. The domain inlet is set $1m$ upstream of the centroid of the building while the outlet is located $3m$ downstream of the origin of the coordinate system [53,54]. The resulting width and height of the domain are 1.5m and 1m respectively.

Two different meshes were constructed for the dDES and the RANS simulations. For the dDES approach, the full domain was discretised with a structured mesh consisting of nearly 7 million cells. The grid is refined close to the ground, around the building and the emitting sources. For the grid sensitivity analysis, two additional grids (one coarser and one finer) were built ($r_h = 1.26$). The coarse one resulted in 3.5 million cells and the fine one in 14 million cells. The relative errors of $U$ and $k$ considering the Coarse-Medium and the Medium-Fine meshes were computed to assess the non-dependence of the result from the grid refinement (Table 2).

<table>
<thead>
<tr>
<th>Refinement</th>
<th>Cells [Millions]</th>
<th>TKE % Error</th>
<th>$U$ % Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coarse Mesh ($f_3$)</td>
<td>3.5</td>
<td>1.65%</td>
<td>1.6%</td>
</tr>
<tr>
<td>Medium Mesh ($f_2$)</td>
<td>7</td>
<td>1.45%</td>
<td>1.45%</td>
</tr>
<tr>
<td>Fine Mesh ($f_1$)</td>
<td>14</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

In addition, the GCI between refinement levels and convergence indexes for the two variables was computed:

$$GCI = \frac{F_s |e|}{r^{p_c} - 1}$$

(11)

where $e$ is the relative error and $p_c = 2$ is the order of convergence. A safety factor $F_s = 1.25$ was assumed [55,56]. The resulting GCIs are: $GCI_{12,TKE} = 3.5\%$, $GCI_{12,U} = 3.4\%$, $GCI_{23,TKE} = 3\%$, and $GCI_{23,U} = 3\%$. As for the RANS approach, due to the symmetry of the model with respect to the plane $y = 0m$, only half of the domain was simulated. A structured mesh consisting of approximately 2.4 million cells was adopted. A grid independence study for the same mesh was carried out by Parente et al. [40].

Table 2: Percentage error of $U$ and $k$ for the three differently refined grids.
Consequently, mesh, numerical model and settings were considered reliable.

Tunnel sides, ground-mounted building, and tunnel top were set as smooth walls, while the tunnel ground was treated as rough wall. The domain inlet was specified as a velocity inlet, setting the turbulence profiles as specified in Table 1. The end of the domain was defined as pressure outlet.

For the dDES simulation, the synthetic turbulence generator was selected for the fluctuating velocity algorithm. The time step size was set to $6e-5s$, ensuring a CFL number below 0.5, with 22 max iterations per each time step. The simulation was run for the time required for the flow to pass through the domain thrice.

**Cedval B1-1 Array of Building** - As equally done in previous studies [57], an array of obstacles is analyzed to validate the proposed $Sc_t$ formulations. As shown in Figure 2 (b), it displays a cluster of $3 \times 7$ obstacles, with the same dimensions of the A1-5 building. The origin of the Cartesian coordinate system was set at the center of the bottom face of the blue building.

The computational domain extends for $6.3m$, $1.5m$ and $1m$ in the $x$, $y$ and $z$ directions respectively with the same orientation adopted for the A1-5 test case. The inlet boundary was set $1m$ upstream the first array of buildings, whereas the outlet was located $4m$ downstream of the last array of buildings. The structured mesh consists of 3.8 million cells. A grid sensitivity analysis was carried out in Longo et al. [39]. The boundary conditions were set analogously to the A1-5 test case.

**CODASC Empty street canyon** - The street canyon from the CODASC dataset [37] involves an empty urban canyon, perpendicular to the inlet velocity, as shown in Figure 2 (c). Its dimensions are specified in Figure 1 (f), with the distance $D$ between the two internal facades (wall A upwind and wall B downwind) equal to the length $L$ and to the height $H$ of the building, $D = L = H = 0.12m$. The origin of the coordinate system is set at the center of the street canyon. The long side of the building measures $L = 1.2m$. The pollutant is emitted simultaneously from 4 line sources, where each of them is $1.42m$ long. The dimensions of the computational domain are $4.92m$, $2m$ and $1m$ in the $x$, $y$ and $z$ directions. The inlet is set $8H$ upstream of the first building and the outlet is located $30H$ downstream of the downwind building. The four source lines were modelled as mass-flow inlet, with $Q_s = 0.02kg/s$. $z_0$ and $u_*$ were specified equal to 0.0033m and 0.535$m/s$ respectively. The remaining boundary conditions were set analogously to the previous test cases. A structured mesh consisting of 3.5 million hexa cells was built. For the grid sensitivity analysis, one additional grid was built, resulting in 2 million cells (coarsening ratio $r_h = 1.2$). The relative errors for velocity and turbulent kinetic energy were estimated equal to 0.35% and 0.45% respectively. Whenever comparing two meshes instead of three, a higher safety factor ($F_S = 3$) is advised [55]. A GCI of 2% was determined for velocity and of 3% for turbulent kinetic energy, with respect to the finest grid.
The simulations were performed in ANSYS Fluent 2019 R3. Second-order schemes were set for the momentum, turbulence quantities and the solute concentration, with a coupled scheme for pressure and velocity. Two tracer gases were selected: sulfur dioxide $SO_2$ for the Cedval test cases A1-5 (single building) and B1-1 (array of building), and sulfur hexafluoride $SF_6$ for the CODASC urban canyon [58].

3. Manifold-Generated Principal component analysis

The objective of PCA is to extract a smaller set of correlated variables, the principal components (PC), that contain most of the variance of the system. These PCs are a linear combination of the original variables and maximize the information contained in the original state-space [32, 59, 60]. Alternatively, Manifold-Generated PCA is a technique that aims at retaining a subset of original variables instead of transforming the entire state-space to a set of principal components. The smaller set of principal variables (PVs) correlates to the detailed model through the usual PCA matrices. The difference lies in working with a reduced set of the original variables rather than a new set of variables, the principal components.

The technique starts with the extraction of high-fidelity data over the entire computational domain and organizing them in a matrix $X$. The size of $X$ equals the number of samples or observations $n$ times the number of conserved variables $Q$. In this case, the conserved variables are a selection of relevant turbulence parameters such as vorticity and strain rate. These parameters will be detailed in the data extraction and analysis part conducted in Section 4.

MG-PCA solves an eigenvalue problem on the covariance matrix of $X$ to obtain the eigenvalues, $L$, and the eigenvectors, $A$:

$$ S = A L A^T. \tag{12} $$

This matrix of eigenvectors can be truncated to a matrix $A_q$ containing only a reduced number $q < Q$ of eigenvectors associated with the highest variance in the system. When projecting the original data on this truncated matrix, the principal components are obtained. These correspond to the most influencing variables of the system (Eq. 13). A reconstruction of the original data can be retrieved by inverting the aforementioned relation:

$$ Z_q = X A_q, \tag{13} $$

$$ \tilde{X}_q = Z_q A_q^T. \tag{14} $$
Manifold-Generated PCA was developed to reconstruct \((Q - q)\) variables of interest from \(q\) observed ones.

The non-principal reconstructed variables are given by the vector \(\tilde{X}(Q - q)\):

\[
\tilde{X}(Q - q) = X(q)BA(Q - q)_q^T,
\]  \hspace{1cm} (15)

where \(B = (A_q^T)^{-1} = X_q^+Z_q\). \(X_q^+\) is the generalized inverse of \(X\), with \(X_q^+ = (X_q^T X_q)^{-1}\).

The present paper explores the use of MG-PCA to find a novel formulation for \(Sc_t\) as a function of a number of imposed principal variables in the dispersion problem. Consequently, the number of principal variables \(q\) equals the number of highly correlated turbulence parameters. In the present case, the variable of interest is \(Sc_t\), meaning that the total number of variables equals \(Q = q + 1\), where \(q\) will be determined using a correlation study as described later in Section 4. Consequently, Eq. 15 can be rewritten as follows: the reconstructed original variable \(\tilde{X}(Q - q)\) is \(Sc_t\), and the principal variables \(X(q)\) are the subset of correlated turbulent parameters. The matrices \(BA(Q - q)_q^T\) are specified according to the MG-PCA method.

4. Results

The main objective is to use a data-driven approach in a multi-fidelity framework to derive a local formulation for \(Sc_t\) based on a set of general turbulence variables. High-fidelity data are obtained through a dDES on the A1-5 single building case [15] with an optimized local formulation of \(Sc_t\) (Eq. 10). Features and correlations are extracted between \(Sc_t\) and the turbulence variables. A first model is derived by calibrating the dDES data to the empirical relation proposed by Reynolds [23]. The outcome of this study leads to a correlation analysis in which a set of highly correlated variables are retrieved. Next, these variables are used with MG-PCA to derive an optimal formulation for \(Sc_t\). The new definition for \(Sc_t\) is verified with additional CFD simulations on different test cases (Figure 1) employing different turbulence models.

The concentration measurements, used for the model validation, are expressed in dimensionless form. For the Cedval test cases:

\[
K = \left(\frac{C_m C_s U_{ref} H^2}{Q_s}\right) Q_s
\]  \hspace{1cm} (16)

and for the CODASC dataset:

\[
K = \left(\frac{x_i U_{ref} H}{Q_l}\right) Q_l
\]  \hspace{1cm} (17)
with $C_m$ and $C_s$ the measured and source tracer concentration in ppm; $U_{ref}$, the reference wind speed in m/s; $H$, the building height (0.125 m for the Cedval cases and 0.12 m for the CODASC one); $Q_s$, total source strength/source flow rate in m$^3$/s; $x_i$, the measured tracer molar fraction; and $Q_l \equiv Q_s/L$, the emission rate of line source m$^2$/s.

4.1. Data from high-fidelity dDES

Data are retrieved from a dDES, employing the ABL inlet conditions. The resulting concentration profiles are displayed in Figure 3. In these profiles, the dDES model coupled with the $Sc_t$ formulation from Eq. 10 is compared to the best-performing RANS approach, namely the ABL turbulence model coupled to the same $Sc_t$ formulation. To verify the effective area of the dDES model where a LES-like approach is applied, it is possible to analyze the dDES TKE dissipation multiplier distribution. The latter demonstrates that in the great majority of the domain a high-fidelity approach is retrieved, as reported in the supplementary material.

The overall prediction of the concentration field is improved with the dDES approach. On top of the building (Figure 3 (a-b)) the DES model is able to properly detect the presence of pollutant. In the same location, similar outcomes resulted from a large eddy simulation run by Gorlé et al. [61]. On the contrary, the RANS approach is not detecting any relevant concentration. The almost total lack of pollutant over the building roof was witnessed with all the tested RANS models. This behavior can be attributed to the intrinsic limitations of the RANS modelling approach. In the downwind locations, the under-prediction of concentration is further reduced by the DES method. As the distance from the pollutant sources increases, the behavior of both the DES and RANS becomes comparable.

To analyze the gain in performance resulting from a variable $Sc_t$ in the DES framework and to prove its robustness, Figure 4 shows the concentration from the dDES runs coupled with a constant $Sc_t = 0.4$, the variable $Sc_t$ formulation from Eq. 10 and the LES results by Gorlé et al. [61]. A small under-prediction of the concentration field can be observed for the DES approach with a constant $Sc_t$. This behavior becomes more accentuated close to the emission location. The comparison demonstrates that the $Sc_t$ plays an equally important role in the DES framework. A similar accuracy was witnessed with respect to the high-fidelity LES results of Gorlé et al. [61].
Figure 3: Comparison of experimental and numerical predictions of non-dimensional concentration for the A1-5 test case at different vertical locations in the symmetry plane. The models employed are: the RANS ABL turbulence model coupled to the previously proposed \(Sc_t\) formulation [10] (red crosses) and the DES model coupled to the same \(Sc_t\) formulation [10] (green triangles).
Figure 4: Comparison of experimental and numerical predictions of non-dimensional concentration for the A1-5 test case at different vertical locations in the symmetry plane. The models employed are: the dDES model coupled to the previously proposed $Sc_t$ formulation [10] (green triangles), the dDES model coupled to a constant $Sc_t = 0.4$ (yellow rhombus) and the LES results by Gorlé et al. [61] employing a variable turbulent dispersion coefficient.
4.2. Correlation study and feature-extraction

Reynolds [23] proposed a semi-empirical formulation, tracing the dependency of $Sc_t$ on $Sc$ and the $\nu_t/\nu$ ratio (Eq. 21). However, there might exist other turbulence variables that show major correlations with $Sc_t$.

The purpose of this investigation is to use the high-fidelity dDES data to verify Reynold’s assumption, and to define a larger set of variables showing major dependencies on $Sc_t$.

To maximize the information contained in the data, the analysis was carried out in planes around the single building of the Cedval A1-5 case. Koeltzsch [24] demonstrated how $Sc_t$ correlates with the height in the boundary layer. Samples were taken at the points corresponding to the experimental locations in the flow field. The test matrix contained $\pm 1000$ observations.

First, the correlation between the $Sc_t$ and a number of flow variables were investigated: the molecular Schmidt number $Sc$, the Reynolds number $Re$, the laminar viscosity $\mu$, the turbulent viscosity $\mu_t$, the turbulent kinetic energy $k$, the turbulent dissipation rate $\epsilon$, the components of the velocity, the laminar Prandtl number, $Pr$, and the Péclet number, $Pe$. The objective is to correlate $Sc_t$ with a set of general parameters that are independent of the used turbulence model. Therefore, the turbulent parameter $C_\mu$ was not included, as it is a model specific variable.

![Figure 5: Correlations between $Sc_t$ and the flow variables analyzing the entire domain (blue) and the $z = 0.01$ m (grey) and $z = 0.035$ m (orange) planes.](image-url)
Figure 5 shows the correlation between $Sc_t$ and the selected flow variables for three different studies: using all the points in the domain, the points in the $z = 0.01m$ plane and the points in the $z = 0.035m$ plane. $Sc$ and $Re$ were raised to the powers 0.7, 0.5 and 0.3. The motivation for raising $Sc$ and $Re$ to an exponent, originates from the empirical formulation of Reynolds where $Sc_t$ is correlated as $Sc_t \approx \exp(Sc^m Re^n)$, with $m$ and $n$ empirically defined constants.

The first conclusion that can be drawn from Figure 5 is that the correlation between $Sc_t$ and other parameters does depend on the boundary layer height. Distinct correlations can be observed in the horizontal planes. However, the variability of the correlations is averaged when studying the entire domain. As a matter of fact, no distinct relation appears between $Sc_t$ and the turbulent variables. This statement changes when focusing on the horizontal planes at the heights $z = 0.01m$ and $z = 0.035m$. Close to the ground, $Sc_t$ correlates with the Reynolds number, the velocity in the x-direction, and the Péclet number,

$$Sc_t \approx \frac{1}{u_x Pe Re}.$$  \hspace{1cm} (18)

When focusing on the $z = 0.035m$ plane, strong correlation can be observed between $Sc_t$ and the Reynolds number raised to a small exponent and the x-velocity,

$$Sc_t \approx \frac{1}{Re^{0.3} u_x}.$$  \hspace{1cm} (19)

Figure 6 compares correlations between $Sc_t$ and various definitions of the Reynolds number: the flow Reynolds number ($Re$), the Reynolds number at the Taylor scale ($Re_\lambda$) and the turbulent Reynolds number ($Re_{turb}$). Higher correlations are found for the xy-plane close to the ground, $z = 0.01m$. $Re_{turb}$ and $Re_\lambda$ show similar contributions, and correlate inversely with respect to the flow Reynolds number. Similar behavior is observed for the $z = 0.035m$ plane. When all the points in the data set are considered, all Reynolds numbers correlate in the same direction.

The conclusions drawn from the correlation study are the following: $Sc_t$ shows different correlations with the flow variables depending on the altitude in the boundary layer, according to previous observations [24, 25]. For the 2 horizontal planes in this investigation, major correlations were found with the Reynolds number, the x-velocity and the Péclet number. When all the points are considered in the study, those specific correlations are averaged and no major correlation can be observed.
Figure 6: Correlations between $Sc_l$ and the Reynolds number in the entire domain (blue) and in the planes $z = 0.01$ m (grey) and $z = 0.035$ m (orange).
4.3. Model calibration based on empirical relations

A first data-driven formulation for $Sc_t$ was derived through the calibration of the dDES data to an empirical law. Reynolds [23] proposed a formulation to retrieve the turbulent Prandtl number in function of the molecular Prandtl and the turbulent viscosity ratio. This empirical law can be extended to $Sc_t$ as follows,

$$Sc_t = C_1 \exp \left[ -C_2 Sc^m \left( \frac{\nu_t}{\nu} \right)^n \right],$$

(20)

with $(\nu_t/\nu)$ the turbulent Reynolds number $Re_t$, and $C_1$, $C_2$, $m$ and $n$ empirically defined constants.

Data for the single building were calibrated to the empirical formulation proposed by Reynolds [23] and verified against the dDES solution with the optimum definition of $Sc_t$ (Eq. 10). Figure 7(a) shows the calibration using the empirical law by Reynolds [23]. Large discrepancies can be observed over the majority of the domain. The large scatter indicates that the combination of the variables proposed by Reynolds [23] does not provide an optimal parametrisation of $Sc_t$. This further suggests that a more appropriate set of turbulence variables could better correlate with a local formulation of $Sc_t$.

The local formulation for $Sc_t$ used in the dDES [15] strictly depends on $C_\mu$, a parameter related to the turbulence model employed. Consequently, $C_\mu$ could appear as a natural choice to replace the turbulent Reynolds number in the empirical law proposed by Reynolds. However, this parameter strongly depends on the turbulence model formulation. Therefore, a dispersion formulation explicitly requiring $C_\mu$ cannot be used in the context of approaches different than the $k-\epsilon$. In this work a calibration to the Reynolds formula is proposed in function of generally available turbulence variables. The Reynolds formula was generalized replacing the turbulent viscosity ratio by the strain invariant $S$ and the vorticity invariant $\Omega$:

$$Sc_t = a \exp \left[ -b Sc^c (d\Omega + eS)f \right].$$

(21)

These invariants are included in the $C_\mu$ definition employed in the simulation but, differently from this model parameter, they can always be retrieved from the flow field. The coefficients $a$, $b$, $c$, $d$, $e$ and $f$ are reported in Table 3.

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.3361</td>
<td>0.6676</td>
<td>1.130</td>
<td>1/3</td>
<td>2/3</td>
<td>0.3668</td>
</tr>
</tbody>
</table>

Table 3: Weighting coefficients obtained in a calibration of $Sc_t$ in function of the vorticity and strain invariants (Eq. 21)

A comparison between the dDES data and the model in Eq. 21 is shown in Figure 7(b). It is clear how the
proposed formulation provides a better parametrisation of $S_{C_t}$ compared to the original formulation.

![Graph](image)

(a) Turbulent Reynolds number  (b) Vorticity and strain rate invariants

Figure 7: Verification of the Reynolds formula against the Cedval A1-5 simulations. The originally imposed correlation with the turbulent Reynolds number (a) is replaced by a linear combination of the strain and vorticity invariants (b).

4.4. Derivation of a generalized $S_{C_t}$ formulation with MG-PCA

A second data-driven formulation for $S_{C_t}$ was derived starting from the dDES data. MG-PCA was used to derive a generalized expression of $S_{C_t}$ based on a number of significant turbulent parameters. These parameters were identified in the correlation study of Section 4.2 and will be imposed as principal variables in the MG-PCA method.

A combination of various flow and turbulence variables were set as principal variables in the MG-PCA according to Eq. 15 with $\tilde{X}(Q - q) = S_{C_t}$. In a first instance, the size of matrix $X$ was $[2,790, 457 \times 11]$, as 11 variables were sampled on almost 3 million locations: the molecular Schmidt number, the Reynolds number, the laminar viscosity, the turbulent viscosity, the turbulent kinetic energy, the turbulent dissipation rate, the components of the velocity, the laminar Prandtl number, and the Péclet number. The goal was to retain a relevant selection of variables to obtain the highest correlation with $S_{C_t}$. The first step was to consider the variables in the Reynolds formulation, i.e. $Re_{turb}$ and $S_c$, and, subsequently, add potentially relevant variables one by one in the MG-PCA model while tracking the model $R^2$ error. The $R^2$ is an indication for the accuracy of the reconstruction between the simulated $S_{C_t}$ and the approximation given by the MG-PCA model.

Starting from the empirical formulation by Reynolds [23], the equation was linearised and the logarithm
of the Reynolds formula was taken. This allows to improve the performances of PCA, which is a multi-
linear approach, and allows to use the $R^2$ as an accuracy metric. To verify the Reynolds formula, the
error metric was computed for the MG-PCA model using the variables specified in the empirical formula,
namely $\log(S_{ct})$, $Sc$, and $Re_{turb}^{0.3}$. A poor correlation of $R^2 = 0.08$ was retrieved. A small improvement was
observed by reducing the exponent of the turbulent Reynolds number from 0.3 to 0.01, the resulting $R^2$
remains unsatisfactory. These results motivate the need for the inclusion of other variables than the ones
suggested by Reynolds in the MG-PCA model.

Considering the variables identified in the correlation study, and adding them to the ones identified by
Reynolds, it is possible to obtain the following optimized formulation:

$$\log(S_{ct}) = a Sc - b Re_{turb}^{c} - d S - e \Omega$$

(22)

with the $a$, $b$, $c$, $d$, $e$ coefficients specified in Table 4.

Table 4: Coefficients for the $S_{ct}$ formulation obtained through MG-PCA (Eq. 22), depending on the molecular Schmidt number, on the turbulent Reynolds, on the strain-rate and vorticity invariants

<table>
<thead>
<tr>
<th></th>
<th>$a$</th>
<th>$b$</th>
<th>$c$</th>
<th>$d$</th>
<th>$e$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.6617</td>
<td>0.8188</td>
<td>0.01</td>
<td>0.0031</td>
<td>0.0329</td>
</tr>
</tbody>
</table>

The final size of matrix $X$ in the MG-PCA model is therefore $[2, 790, 457 \times 5]$. The constraint of adopting the
same turbulence model as the one employed in this study is solved. This is a relevant aspect, considering
that this $S_{ct}$ formulation can be potentially employed with other turbulence models (LES, DES, RANS:
standard-realizable-RNG $k - \epsilon$, $k - \omega$, RSM e.g.), without invalidating its local variability. The latter will be
verified in Section 4.5.
4.5. Verification on the selected test cases

The resulting $Sc_t$ formulations trained with the DES data are tested and validated using RANS approaches over the selected test cases: Cedval A1-5 single building, Cedval B1-1 array of buildings and CODASC empty street canyon (Figure 2). They are compared against the constant $Sc_t$, the formulation proposed by Gorlé et al. [29], and the one by Longo et al. [15]. Additionally, different RANS turbulence closures were tested ranging from the standard ones to the optimised closure for ABL flows.

The main properties of the analyzed pollutants are briefly listed in the supplementary material, while the $Sc_t$ formulations used in this study are presented in Table 5.

Table 5: List of Schmidt formulations and their main parameters employed and compared in this study

<table>
<thead>
<tr>
<th>Author</th>
<th>Formulation</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>$Sc_t = 0.4$</td>
<td>$-$</td>
</tr>
<tr>
<td>Gorlé et al. [29]</td>
<td>$Sc_t = \frac{9}{8} C_\mu C_0$</td>
<td>$C_0 = \frac{C_{\infty}}{1 + 7.5 C_{\infty} Re_\lambda^{-1}}$</td>
</tr>
<tr>
<td>Longo et al. [15]</td>
<td>$Sc_t = \frac{2 C_\mu}{C_0 C^2}$</td>
<td>$C = 0.35, C_0 = 2$</td>
</tr>
<tr>
<td>Present work (Eq. [21])</td>
<td>$Sc_t = a \left[ -b Sc^e (d \Omega + e S)^f \right]$</td>
<td>$a, b, c, d, e, f$</td>
</tr>
<tr>
<td>Present work (Eq. [22])</td>
<td>$Sc_t = \exp[a Sc + b Re_{turb}^c + d S + e \Omega]$</td>
<td>$a, b, c, d, e$</td>
</tr>
</tbody>
</table>
4.5.1. Cedval A1-5 Single Building

The first set of simulations of the Cedval A1-5 is meant to compare 4 different configurations against the experimental data. This full set of data is plotted in Figure 8:

- standard $k - \epsilon$ model, $Sc_t = 0.4$
- ABL model, $Sc_t$ proposed by Gorlé et al. [29]
- ABL model, $Sc_t$ based on MG-PCA (Eq. 22)
- ABL model, revisited Reynolds [23] $Sc_t$ formulation (Eq. 21).

All the models displaying a variable $Sc_t$ definition show an increased accuracy with respect to a constant $Sc_t$. As expected, the proposed data-driven $Sc_t$ formulations are able to replicate the performance of the optimal $Sc_t$ (Eq. 10). They show enhanced accuracy with respect to the standard model and a further improvement with respect to the model proposed by Gorlé et al [29].

In Figure 9 (a-c), the standard $k - \epsilon$ model is employed with the $Sc_t$ formulation derived with MG-PCA. Predictably, the variability of the latter is conserved also when an approach displaying a constant $C_\mu$ is adopted and permits an increase in the accuracy with respect to the application of a constant $Sc_t$. Similar observations can be drawn from Figure 9 (d-f), where the MG-PCA approach is applied coupled to the realizable $k - \epsilon$ model and the Reynolds formulation with the Generalized $k - \omega$ model. Both models show similar performance with respect to the optimum approach, namely the one employing the $Sc_t$ proposed by Longo et al. [15]. Additional results for this set of simulations can be found in the supplementary material.
Figure 8: Comparison of experimental and numerical predictions of non-dimensional concentration for the A1-5 test case at different vertical locations, in the symmetry plane (a-e) and in the \( y = -0.06 \) m plane (f), as displayed in Figure 2 (a). The models used are: the standard \( k-\varepsilon \) with \( Sc_t = 0.4 \) (green dashed line), the ABL turbulence model with the \( Sc_t \) by Gorlè et al. [29] (light blue dotted line), the ABL turbulence model coupled to the previously proposed \( Sc_t \) formulation [10] (red dashed line) and to the one based on Reynolds formulation [21] (orange crosses).
Figure 9: Comparison of experimental and numerical predictions of non-dimensional concentration for the A1-5 test case at different vertical locations, in the symmetry plane, as displayed in Figure 2 (a). On top (a-b-c), the models employed are: the standard $k-\varepsilon$ with $Sc_t = 0.4$ (green dashed line), with the MG-PCA formulation (grey broken line), and the ABL turbulence model with the optimum $Sc_t$ formulation (red crosses). On bottom (d-e-f), the models employed include the MG-PCA formulation coupled to the realizable $k-\varepsilon$ model (blue dotted line), the Reynolds formulation coupled to GEKO (violet broken line) and the ABL turbulence model coupled to the previously proposed $Sc_t$ formulation (red crosses)
4.5.2. Cedval B1-1 Array of buildings

An array of buildings is considered for validating the proposed methodologies. Experimental profiles are localized, available for the horizontal plane at 7.5mm from the ground (Figure 2(b)).

In Figure 10, the standard $k-\epsilon$ model with $Sc_t = 0.4$ is compared with the ABL turbulence model coupled with the $Sc_t$ by Gorlé et al. [29], the $Sc_t$ resulting from the MG-PCA (Eq. 22) and from the verification of Reynolds formulation (Eq. 21). In the considered locations, the standard $k-\epsilon$ tends to overestimate the concentration field, especially in the downwind locations (Figures 10 (b-f)). When the ABL model is employed, more accurate predictions are witnessed. Specifying $Sc_t$ according to Gorlé et al. [29] leads to a good agreement between predictions and experiments, with respect to the standard methodology. Good agreement with experimental data is witnessed when using the MG-PCA based $Sc_t$ formulation (red broken line) and from the empirical Reynolds formulation (orange crosses). The over-prediction of the concentration is further limited, resulting in slight discrepancies at few locations (Figure 10 (a)(c)) with an average deviation of 9% and 6%.

Additional results can be found in the supplementary material.
Figure 10: Comparison of experimental and numerical predictions of non-dimensional concentration for the B1-1 test case at different horizontal locations, as displayed in Figure 2 (b). The models used are: the standard $k-\epsilon$ with $Sc_t = 0.4$ (green dashed line), the ABL turbulence model with the $Sc_t$ by Gorlé et al. [29] (light blue dotted line), the ABL turbulence model coupled to the $Sc_t$ formulation based on MG-PCA [22] (red dashed line) and to the one based on Reynolds [21] (orange crosses).
Finally, the proposed $Sc_t$ formulations are validated on an empty street canyon case. The performance of the variable formulations are compared to the one of the standard $k-\varepsilon$ model with $Sc_t = 0.4$, at three different heights on the internal facades of both the upwind building (building A, measurement plane located at $x = -0.055m$) and the downwind building (building B, measurement plane located at $x = 0.055m$), as shown in Figure 2 (c). The dimensionless concentration profiles are presented in Figure 11. Due to the symmetry of the problem, only half of the profiles ($-0.6m < y < 0m$) are displayed.

It can be observed that the standard methodology tends to overestimate the pollutant concentration. This is especially true close to the ground, where the pollutant is accumulated. One exception to this trend is witnessed at location $z = 0.10m$ on the facade of building B (Figure 11 (f)), where the standard methodology predicts the concentration field accurately. This over-prediction is magnified at the central locations of the wall ($-0.4m < x < 0.4m$).

Even if a certain discrepancy with respect to the experimental data is still observed, more accuracy is shown by the approaches employing the variable $Sc_t$ formulations. Considering the $Sc_t$ proposed in this study, the best agreement with experimental data is witnessed close to the ground. Accurate performance is also shown by the formulation by Gorlé et al. [29]. Once again, the major over-prediction is located at the center of the facades. A similar trend was experienced in previous studies on the same case study [62].

The local variability of the proposed $Sc_t$ based on MG-PCA [22] is demonstrated in the contour plots in Figure 12. $Sc_t$ is ranging according to the local turbulence level, directly depending on $Sc$, $S$ and $\Omega$.

Additional results can be found in the supplementary material.
Figure 11: Comparison of experimental and numerical predictions of non-dimensional concentration for the empty street canyon test case at different horizontal axial locations (Figure 2), using the standard $k - \varepsilon$ with $Sc_t = 0.4$ (green dashed line), the ABL turbulence model with the MG-PCA $Sc_t$ (red broken line) and the ABL turbulence model coupled to the re-proposed Reynolds formulation depending on vorticity and strain rate (blue dotted line).
Figure 12: Contour plots of $S_{C_t}$ derived with MG-PCA (equation 22) for the 90° oriented empty street canyon test case from the CODASC dataset in the horizontal $z = 0.035m$ plane (a) and in the vertical $y = 0m$ plane (b). The turbulence approach employed is the ABL RANS model.
5. Conclusion

A comprehensive multi-fidelity framework was presented to derive a local formulation of $S_{CT}$ starting from high-fidelity data. dDES were performed on the Cedval A1-5 single building test case to produce a reliable database. These data were used in a correlation study to identify a set of flow and turbulence parameters that identify with $S_{CT}$. A first data-driven model was proposed by calibrating the dDES samples to a semi-empirical formulation proposed by Reynolds. The original formulation by Reynolds was improved by replacing the initial parameters in the model, namely the molecular Schmidt number and the turbulent Reynolds number, by a linear combination of the strain and vorticity invariants. A second formulation was derived by correlating a set of model variables, i.e. the molecular Schmidt number, the turbulent Reynolds number, the strain rate invariant and the vorticity invariant, using Manifold Generated-PCA. In total, 2 new variable turbulent Schmidt formulations were proposed in this study. Both formulations were verified with lower-fidelity RANS simulations on the Cedval A1-5 single building test case, and two additional test cases, namely the Cedval B1-1 array of building, and an empty canyon of the CODASC database. Moreover, the compatibility of the model was verified in combination with various turbulence approaches. The two data-driven models were compared with simulations using a constant $S_{CT}$, and several local formulations proposed previously in the literature. The resulting $S_{CT}$ numbers, based on the calibration of the Reynolds-like formula and on Manifold Generated-PCA, capture most of the dependencies observed for the $S_{CT}$ in literature, fall in the expected range of atmospheric $S_{CT}$ values and are able to improve substantially the prediction of the concentration field. In this regard, the prediction of the concentration field was improved up to 100%, when the variable Schmidt number formulations were applied with both the ABL and standard turbulence models.

Future work will further investigate the use of turbulence models that account for the anisotropic character of the turbulent dispersion. Moreover non-neutral atmospheric stability conditions and more challenging test cases in terms of orography will be analyzed.


[37] University of Karlsruhe. Codasc stands for “concentration data of street canyons”. codasc data is from the laboratory of building- and environmental aerodynamics at the institute for hydromechanics (ihf) at the university of karlsruhe/germany. [https://www.windforschung.de/CODASC.htm](https://www.windforschung.de/CODASC.htm).

[38] W. Rodi. Comparison of les and rans calculations of the flow around bluff bodies. *Journal of wind engineering and industrial


