The THESAN project: properties of the intergalactic medium and its connection to reionization-era galaxies

E. Garaldi,[1]⋆ R. Kannan,[2]† A. Smith,[3]‡ V. Springel,[1] R. Pakmor,[1] M. Vogelsberger[3] and L. Hernquist[2]

[1]*Max-Planck Institute for Astrophysics, Karl-Schwarzschild-Str. 1, D-85741 Garching, Germany*
[2]*Center for Astrophysics | Harvard & Smithsonian, 60 Garden Street, Cambridge, MA 02138, USA*
[3]*Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

ABSTRACT

The high-redshift intergalactic medium (IGM) and the primeval galaxy population are rapidly becoming the new frontier of extragalactic astronomy. We investigate the IGM properties and their connection to galaxies at $z \geq 5.5$ under different assumptions for the ionizing photon escape and the nature of dark matter, employing our novel THESAN radiation-hydrodynamical simulation suite, designed to provide a comprehensive picture of the emergence of galaxies in a full reionization context. Our simulations have realistic 'late' reionization histories, match available constraints on global IGM properties and reproduce the recently-observed rapid evolution of the mean free path of ionizing photons. We additionally examine high-z Lyman-α transmission. The optical depth evolution is consistent with data, and its distribution suggests an even-later reionization than simulated, although with a strong sensitivity to the source model. We show that the effects of these two unknowns can be disentangled by characterising the spectral shape and separation of Lyman-α transmission regions, opening up the possibility to observationally constrain both. For the first time in simulations, THESAN reproduces the modulation of the Lyman-α flux as a function of galaxy distance, demonstrating the power of coupling a realistic galaxy formation model with proper radiation-hydrodynamics. We find this feature to be extremely sensitive on the timing of reionization, while being relatively insensitive to the source model. Overall, THESAN produces a realistic IGM and galaxy population, providing a robust framework for future analysis of the high-z Universe.

Key words: galaxies: high-redshift – cosmology: dark ages, reionization, first stars – radiative transfer – methods: numerical

1 INTRODUCTION

Within a billion years of the Big Bang, the gas content of the Universe changed its ionization state twice: it became neutral at the recombination epoch as a consequence of the cooling provided by the expansion of the Universe (producing the Cosmic Microwave Background, or CMB), and it was later ionized again (most likely) by the first generations of galaxies during a period known as the Epoch of Reionization (EoR, e.g. Shapiro & Giroux 1987; Loeb & Barkana 2001; Haiman 2016).

Reionization is the least understood period of time in the history of the Universe, because of extreme observational difficulties in probing this remote epoch and formidable challenges to its theoretical investigation. However, understanding its unfolding is essential, as it constitutes the evolutionary link between the high-redshift Universe and the cosmic structures today. As such, it plays an important role in the origin of galaxies and provides a test-bed for evaluating our knowledge – mostly gained from local observations – of the relevant physical processes in galaxy formation.

From decades of investigations, a relatively-coherent picture of the EoR has been built. The modeling of anisotropies in the CMB (Planck Collaboration et al. 2020), the evolution of the Lyman-α (hereafter Lyα) forest opacity (Becker et al. 2001; Fan et al. 2006; Becker et al. 2015), and of the number density of detected Lyα-emitting systems (Ota et al. 2010; Pentericci et al. 2011; Mason et al. 2018), the damping wings of high-redshift quasars (e.g. Schroeder et al. 2013; Greig et al. 2017; Davies et al. 2018), and the statistics of dark pixels in quasar absorption spectra (e.g. McGreer et al. 2011) have allowed us to constrain the EoR to the redshift range $5.5 \lesssim z \lesssim 10$ and its tail-end (where individual ionized regions coalesce) to occur at $5 \lesssim z \lesssim 6$. It is now mostly accepted that this process was mainly powered by a swarm of small (i.e. with stellar mass $M_{\rm star} \lesssim 10^8\,{\rm M}_\odot$), star-forming galaxies with relatively high escape fractions of ionizing photons (Madau et al. 1999; Gnedin 2000; Haardt & Madau 2012), although in recent times some evidence against this picture has been presented (Naidu et al. 2020). Active galactic nuclei (AGN, Haardt & Madau 1996; Kulkarni et al. 2019b) are thought to have played only a minor role at redshifts $z \gtrsim 5$ (but see Madau & Haardt 2015; Garaldi et al. 2019a).

Observationally confirming this picture, however, requires understanding the evolution of the intergalactic medium (IGM) between galaxies as well as the formation of the first structures. This task is arduous. For instance, to date only $\mathcal{O}(1000)$ galaxy candidates have been discovered at $6 \lesssim z \lesssim 8$ (and much fewer at higher redshift, see e.g. Bouwens et al. 2015; Livermore et al. 2017; Atek et al. 2018),

many of which do not yet have spectroscopic characterisations. Nevertheless, the formation of primeval galaxies above redshift $z \gtrsim 6$ is rapidly becoming the new frontier in the study of both cosmic reionization and galaxy formation, as evidenced by recent observational campaigns that have started to constrain the *population* of galaxies at the end of the EoR, e.g. REBELS (Bouwens et al. 2021) and ALPINE (Le Fèvre et al. 2020). These are (or will soon be) flanked by many instruments providing us with a multi-faceted picture of the emergence of primeval galaxies in the context of the EoR. Among them, the Atacama Large Millimiter Array (ALMA), the *James Webb Spaces Telescope* (JWST), the Cerro Chajnantor Atacama Telescope-prime(CCAT-p), and the Spectro-Photometer for the history of the Universe, Epoch of Reionization and Ices Explorer (SPHEREx) focus on the properties of high-z galaxies, while the Low-Frequency Array (LOFAR), the Murchison Widefield Array (MWA), the Hydrogen Epoch of Reionization Array (HERA), and the Square Kilometer Array (SKA) investigate instead the distribution of neutral hydrogen before and during the EoR, either statistically or tomographically.

These new-generation instruments are designed to assist the traditional means of investigations of the EoR, which mainly relied on the Lyα transition in neutral hydrogen atoms, the most prominent and ubiquitous absorption line in quasar spectra. However, because of its large absorption cross section, even a modest hydrogen neutral fraction of $x_{\mathrm{H\,I}} \sim 10^{-4}$ completely suppresses the incoming flux. This has limited the reach of Lyα absorption studies to the post-reionization Universe until recent years, when technological advancements (in particular better spectral resolution and sensitivity) coupled with the discovery of QSOs at higher and higher redshift have allowed us to probe the end phases of the EoR (e.g. Barnett et al. 2017; Kakiichi et al. 2018; Meyer et al. 2019; Bosman et al. 2018; Eilers et al. 2018; Yang et al. 2020).

These advancements provided new fuel for refined theoretical investigations. For instance, Kulkarni et al. (2019a) and Keating et al. (2020) combined these observations with post-processing radiation transport (RT) simulations to provide evidence in support of a completion of the reionization process as late as $z \sim 5.2$ (now known as a 'late' reionization model), which appears necessary to match the distribution of optical depth at $z \lesssim 6$. Gnedin et al. (2017), Garaldi et al. (2019b), and Gaikwad et al. (2020) all characterised the properties of the Lyα transmission regions at $z \lesssim 6.5$ in order to establish the fidelity of current state-of-the-art simulations and connect the geometrical properties of such transmission regions to the underlying physical conditions of the IGM between galaxies. Notably, Garaldi et al. (2019b) demonstrated that even state-of-the-art radiation-hydrodynamical simulations are unable to match the observed connection between Lyα transmission in QSO spectra and the galaxy population at $z \sim 5.5$.

The aforementioned work represents an example of how the connection between cosmic reionization and the galaxy population is one of the most challenging and promising grounds for the investigation of the high-z Universe. In fact, the convergence of the fields of cosmic reionization and galaxy formation poses a fierce new challenge; i.e. to simultaneously understand the properties of the intergalactic medium (IGM) on large scales and of high-redshift galaxies on small scales, as well as – crucially – their connection in a unique and consistent picture.

In an attempt to improve our understanding of the EoR, a number of different numerical approaches have been developed. However, numerical studies of the EoR are among the most challenging in the landscape of numerical astrophysics. The reason lies in the requirement to simultaneously: (i) cover large volumes, in order to provide a statistically-representative picture of the reionization process; (ii)

resolve (sub-)galactic scales, where processes responsible for the production of ionizing photons occur, where the self-shielded gas acting as photons sink resides, and where galactic observables are produced; and (iii) include an accurate treatment of RT. The gargantuan computational cost of simulations with these characteristics has forced most of the efforts to date to employ simplifying approximations. In fact, while fully-coupled radiation-hydrodynamics simulations have become possible in recent years, they are often limited to small volumes or individual galaxies (e.g. Pawlik et al. 2017; Rosdahl et al. 2018; Pallottini et al. 2017; Katz et al. 2019; Xu et al. 2016; Trebitsch et al. 2020a), or coarse resolutions (e.g. Ocvirk et al. 2016, 2020), although notable exceptions exist (Gnedin 2014). A common issue of these simulations is the fact that they are too expensive to be run to $z = 0$, and therefore their galaxy formation models are only calibrated and tested against $z \gtrsim 5$ (scarce and of limited quality) observations. This renders the fidelity of the simulated physics uncertain. In fact, when some of these models have been evolved to lower redshift, they produced galaxy properties incompatible with observed ones (e.g. Tacchella et al. 2018; Trebitsch et al. 2020b; Mitchell et al. 2021).

In order to achieve a comprehensive and statistically-significant view of the high-z Universe and bypass the aforementioned problems, we have developed the THESAN simulation suite, which we introduce in this and two companion papers. In particular, Kannan et al. (2021, hereafter Paper I) presents an overview of the simulations, while Smith et al. (2021, hereafter Paper II) describes results concerning the Lyα transmission from individual haloes. This paper focuses on the properties of the diffuse IGM and its Lyα forest in the THESAN suite, including a thorough comparison with available observations, shedding new light on the reionization process and the physics of high-redshift Universe. THESAN provides a unique combination of large simulated volumes, high numerical resolution, a wealth of physical processes simulated in a self-consistent manner (including RT and cosmic dust), exploration of high-redshift physics (including escape fraction and dark matter models), and a realistic galaxy formation model at *all* redshifts. We intend that this suite will provide a firm ground upon which future investigations can be built. For this reason, we will make the data public in the near future and welcome proposals for collaborative work.

The paper is organised as follows. In Section 2 we present a brief overview of the simulation suite and details of the forward modelling performed in this study. In Section 3 we describe results concerning the global properties of the IGM. In Section 4 we focus on the high-z Lyα transition. In Section 5 we investigate the connection between the high-z IGM and coeval galaxy population. Finally, we provide a summary and concluding remarks in Section 6.

2 METHODS

In this Section we summarise the main features of the THESAN simulations, which are more thoroughly described in Paper I. These represent the backbone of the analysis presented here, as they provide the physical state of the simulated universe. In order to fruitfully compare simulated and observed quantities, we post-process our simulations to produce synthetic observations that are as similar as possible to real data. We describe how this is achieved at the end of this Section.

2.1 The THESAN simulations

In this work we employ the THESAN simulation suite of radiation-magneto-hydrodynamical simulations run with the moving-mesh hy-

Table 1. A summary of the main properties of the THESAN simulations employed in this paper. From left to right the columns report the name of the simulation, box size, initial particle number, mass of the dark matter particles and gas cells, the (minimum) softening length of (gas) star and dark matter particles, minimum cell size at $z = 5.5$, the final redshift, the escape fraction of ionizing photons from the birth cloud (if applicable) and a short description of the simulation. A complete version of this Table can be found in Paper I.

Name	$L_{\rm box}$ [cMpc]	$N_{\rm particles}$	$m_{\rm DM}$ [M$_\odot$]	$m_{\rm gas}$ [M$_\odot$]	ϵ [ckpc]	$r_{\rm cell}^{\rm min}$ [pc]	$z_{\rm end}$	$f_{\rm esc}$	Description
THESAN-1	95.5	2×2100^3	3.12×10^6	5.82×10^5	2.2	~ 10	5.5	0.37	fiducial model (RMHD + TNG + dust)
THESAN-2	95.5	2×1050^3	2.49×10^7	4.66×10^6	4.1	~ 35	5.5	0.37	fiducial model
THESAN-WC-2	95.5	2×1050^3	2.49×10^7	4.66×10^6	4.1	~ 35	5.5	0.43	fiducial model + weak convergence of $x_{\rm HI}(z)$
THESAN-HIGH-2	95.5	2×1050^3	2.49×10^7	4.66×10^6	4.1	~ 35	5.5	0.8	fiducial model + $f_{\rm esc} = 0$ if $M_{\rm halo} \geq 10^{10} h^{-1}$ M$_\odot$
THESAN-LOW-2	95.5	2×1050^3	2.49×10^7	4.66×10^6	4.1	~ 35	5.5	0.95	fiducial model + $f_{\rm esc} = 0$ if $M_{\rm halo} < 10^{10} h^{-1}$ M$_\odot$
THESAN-SDAO-2	95.5	2×1050^3	2.49×10^7	4.66×10^6	4.1	~ 35	5.5	0.55	fiducial model + sDAO dark matter model
THESAN-TNG-2	95.5	2×1050^3	2.49×10^7	4.66×10^6	4.1	~ 35	5.5	-	MHD + TNG model (original TNG)

drodynamics code AREPO (Springel 2010; Weinberger et al. 2020)[1]. The equations of radiation-magneto-hydrodynamics are solved on a mesh corresponding to the Voronoi tessellation of mesh-generating points that approximately follow the gas flow. The computation of gravity relies on a hybrid Tree-PM approach, which separately computes long-range (via a particle mesh approach) and short-range (through a hierarchical oct-tree, Barnes & Hut 1986) forces. In addition, we employ a hierarchical time integration, allowing an efficient management of the deep time bin hierarchy, and a randomization of the node centres at each domain decomposition (described in the GADGET4 paper, Springel et al. 2021) that suppresses force errors at the boundary of large nodes.

2.1.1 Galaxy formation model

The THESAN simulations were designed to simultaneously capture the assembly of primeval galaxies and their coupling to reionization, while maximising their physical fidelity. For this reason, we choose to employ a sub-resolution physics model that has been extensively investigated at low redshift, namely the IllustrisTNG model (Nelson et al. 2018a; Pillepich et al. 2018b; Springel et al. 2018; Marinacci et al. 2018; Naiman et al. 2018; Nelson et al. 2019b; Pillepich et al. 2019). (Details of the TNG model can be found in Weinberger et al. 2017, Pillepich et al. 2018a, Nelson et al. 2019a. The model is an improvement over the Illustris one, described in Vogelsberger et al. 2014a and Vogelsberger et al. 2014b. We note here that this model employs a tabulated spatially-uniform UV background, that we replace with self-consistent RT as described in the next Section.) This choice ensures that, although the THESAN simulations are not evolved below $z = 5.5$, their physical model and the galaxy population they produce can be trusted throughout the entire history of the Universe. We present supporting evidence for this claim in Paper I. In particular, we show that THESAN is able to produce a realistic stellar-to-halo-mass relation (figure 9 of Paper I), stellar mass function (figure 10 of Paper I), UV luminosity function (figure 11 of Paper I), star-formation-rate-density evolution (figure 12 of Paper I), and mass-metallicity relation (figure 14 of Paper I). In addition, Wu et al. (2019) demonstrated in a very similar configuration that replacing the Faucher-Giguère et al. (2009) UV background (UVBG) with self-consistent RT does not significantly alter the main properties of galaxies, while having a large impact on the properties of the IGM.

Finally, we check in Appendix A that this holds true in our simulations as well by employing the THESAN-2 and THESAN-TNG-2 runs. All these results make us confident that the successes of the IllustrisTNG model will be retained in the THESAN model at all redshifts. Additionally, employing the IllustrisTNG model *as is* (with the exception of the UVBG being replaced by the self-consistent RT) effectively limits the free parameters to a single number, the (unresolved) stellar escape fraction $f_{\rm esc}$ (which quantifies the sub-resolution absorption and is not present in the IllustrisTNG model as it lacks RT), which we calibrate against the reionization history of the Universe. In fact, all other physical parameters have been tested against low-redshift observables during the development of the IllustrisTNG simulations. We note here that, among other physical processes, the IllustrisTNG model and therefore the THESAN simulations as well include black holes, and magnetic fields originating from a uniform, cosmological seed field of initial value $B_{\rm seed} = 10^{-14}$ G. The former are created placing a seed black hole of mass $M_{\rm seed} = 8 \times 10^5 h^{-1}$ M$_\odot$ at the center of every FOF group that exceeds a mass threshold of $M_{\rm thr} = 5 \times 10^{10} h^{-1}$ M$_\odot$. They then accrete material from neighbouring cells using a Bondi-Hoyle prescription (Bondi & Hoyle 1944; Bondi 1952) and release energy into the surrounding gas via two feedback modes (dubbed quasar and radio mode, see Weinberger et al. 2017). We refer the reader to Paper I for an overview of the black hole population in the THESAN simulations, and remark here that they globally contribute less than 1% of the hydrogen-ionizing photons.

2.1.2 Self-consistent radiative transfer

In order to self-consistently solve the RT equations, we employ the AREPO-RT (Kannan et al. 2019) extension of the AREPO code, which solves the first two moments of the RT equation, complemented with the M1 closure relation (Levermore 1984). Because of computational limitations, we decide to follow only the UV part of the radiation spectrum and discretise the photon energy into bins defined by the following energy thresholds: $[13.6, 24.6, 54.4, \infty)$ eV, chosen to braket the ionization energies of hydrogen and helium. Each resolution element tracks, for each bin, the comoving photon number density and flux. In order to partially compensate the loss of resolution in the frequency sampling, which may introduce errors in the temperature and ionization state of the gas (see e.g. Mirocha et al. 2012), we assume the radiation follows, within each bin, the spectrum of a 2 Myr old, quarter-solar metallicity stellar population with amplitude given by the photon number density of the cell itself. The values of the integral average over the bin of the H I, He I and He II
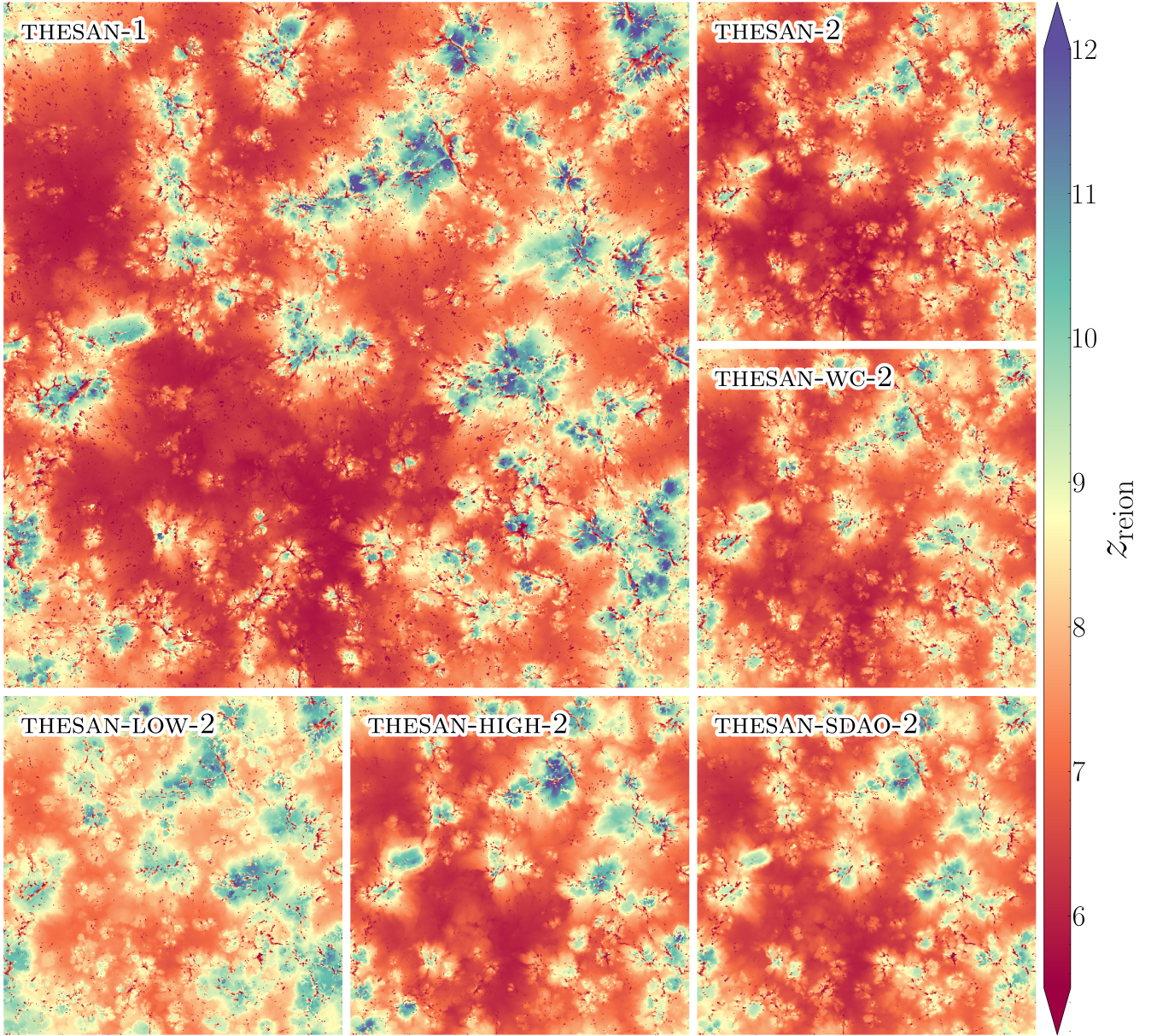

Figure 1. Reionization redshift, defined as the latest time at which the ionized fraction of the resolution element crosses the threshold value 0.99, for each resolution element in a slice through a selection of the THESAN simulations. The different escape fraction and dark matter models produce visible differences in the reionization redshift distribution. Encouragingly, the fiducial and weak convergence runs show very few differences.

cross section (σ_X), energy injected per unit photon ($\mathcal{E}_X$), and photon energy (e) are reported in Table 1 of Paper I. The stellar spectra are computed as a function of mass, metallicity and age of each stellar particle using the Binary Population and Spectral Synthesis models v2.2.1 (BPASS, Eldridge et al. 2017; Stanway & Eldridge 2018), assuming a Chabrier IMF (Chabrier 2003). Additionally, we include radiation from quasars by converting the black hole accretion rate computed by the code into a quasar luminosity (assuming a radiative efficiency of $\eta_{\rm rad} = 0.2$) and assuming the spectral shape from Lusso et al. (2015) with unit escape fraction. Finally, we employ a reduced speed of light (RSLA) approximation, with an effective value of the speed of light $\tilde{c} = 0.2c$. Recently, there has been a debate in the literature on whether employing a RSLA produces appreciable effects

on the reionization of the Universe (Bauer et al. 2015; Gnedin 2016; Deparis et al. 2019; Ocvirk et al. 2019). We have carefully chosen the value of the (reduced) speed of light to prevent any sizeable effect on the IGM (as measured by the reionization front propagation speed and reionization history, see Appendix A of Paper I). We are therefore confident that the reionization properties in THESAN are not affected by this approximation. Additionally, Gnedin (2016) showed that the RSLA approximation does not impact star-formation properties of galaxies for our employed value of $\tilde{c}$, supporting our expectation that the galaxy properties are robust even when employing the RSLA. Overall, we believe that the use of a (mild) RSLA does not have appreciable effects on the results presented in this paper.

The RT equations are coupled to a non-equilibrium solver that
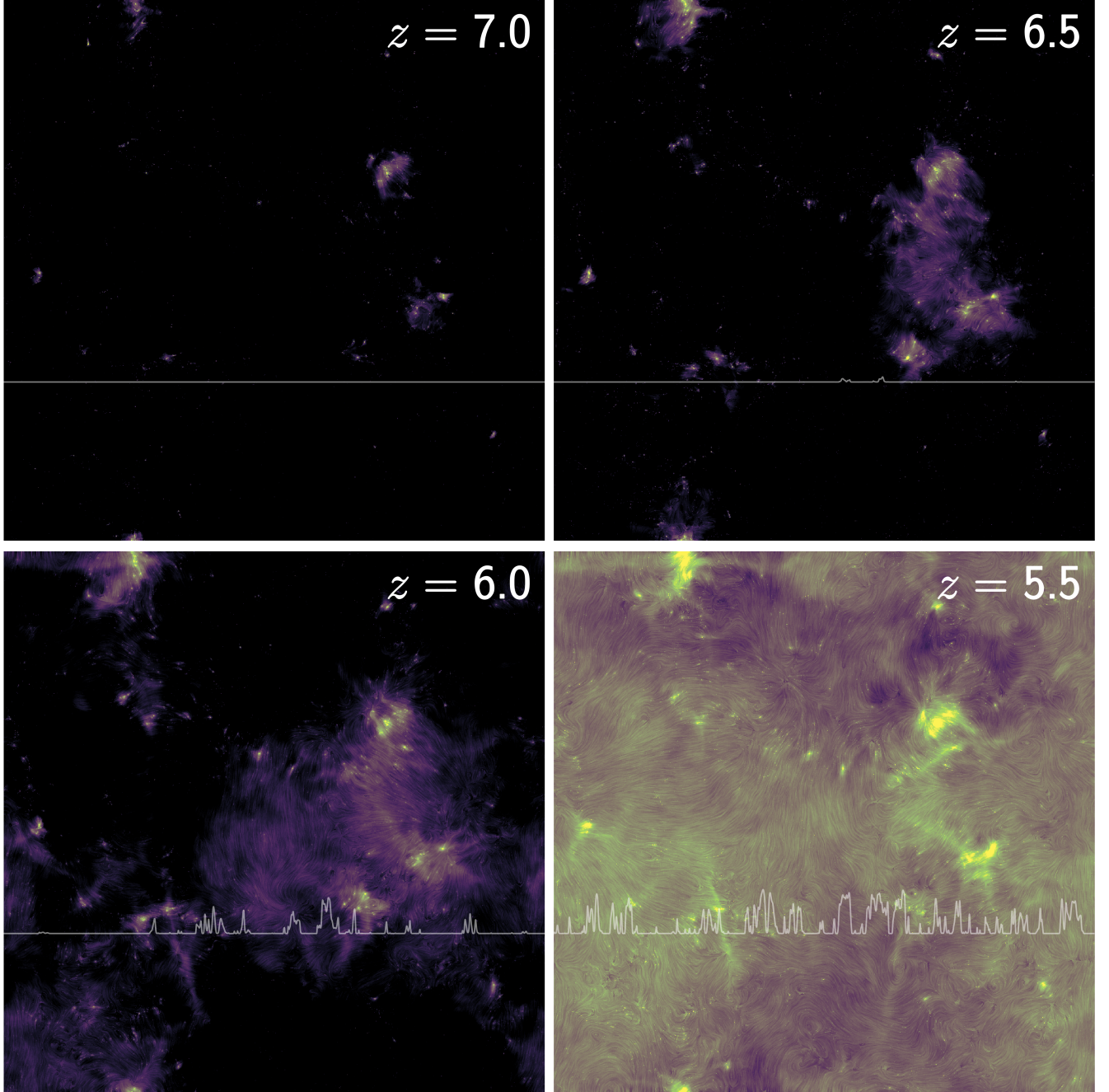
Figure 2. Visualization of a slice (of thickness equal to 20% of the box size) through THESAN-1. Panels show the line integral convolution of the photon density (providing the background texture) and the photon flux along the image plane (providing the vector field). In each panel, the redshift is indicated in the top right corner. Galaxies and galaxy groups act as sources of photons, which travel approximately linearly until they encounter radiation sinks. Notice that the color scale employed is identical in all panels. The change in the overall hue of the panels reflects the rapid evolution of the ionising photon field in the latest phases of reionization. Additionally, we show in each panel the Lyα transmitted flux in a single sightline traversing the slice, showing the rapid build-up of transmitted flux in ionised regions at $z \lesssim 6$.

accurately computes the ionization state of hydrogen and helium, as well as the temperature change due to photo-ionization, metal cooling and Compton cooling. Finally, in order to provide a more faithful ground of comparison with observations of high-z dust, we have included in the THESAN simulation the dust model developed by McKinnon et al. (2016, 2017).

The inclusion of a realistic reionization history implies that galax-

ies in THESAN are subject to a spatially-varying UV radiation field. This is expected to affect galaxies by means of radiative feedback on the gas (e.g. radiation pressure, photo-ionization and photo-heating) in a different way than the combination of spatially-uniform UVBG and approximate self-shielding corrections that is usually employed in cosmological hydrodynamical simulations (including in the Illus-trisTNG model). As discussed, however, THESAN describes the ISM

through an effective equation of state (eEOS, originally developed in Springel & Hernquist 2003). In this model, the ISM gas is assumed to be composed of a volume-filling hot and diffuse component in pressure equilibrium with a cold and dense phase, where stars are formed. From the gas density, its temperature, star-formation rate and other properties can be inferred. The advantage of such a treatment is that the global properties of the ISM gas are captured without the need to resolve physical processes on small scales, making it possible to efficiently run accurate galaxy simulations of cosmological volumes (see e.g. the IllustrisTNG simulations, or the Auriga suite, Grand et al. 2017). However, this simplicity comes at the cost of an ISM composition that is assumed a priori and involves a tight relation between gas density and temperature. This rigidity means that feedback processes can affect the gas in only two ways, namely they either: (i) heat the gas enough to prevent it from cooling back to the eEOS temperature within a single time step of the simulation, hence preventing star formation for the time step, or (ii) they affect the gas *outside* the galaxy, altering its accretion onto it. The latter occurs through e.g. photo-heating of the diffuse gas inside and outside the halo (e.g. Okamoto et al. 2008). Therefore, simulating the ISM using an eEOS model partially hides the effects of radiative feedback on the small-scale structure of the ISM. Nevertheless, this approach captures the indirect effects of photo-heating, as shown by Wu et al. (2019) and discussed in the previous section as well as in Appendix A. Additionally, galactic properties are well captured, as shown by the large body of simulations employing this approach and reproducing many observations.

2.1.3 Simulation design

All THESAN simulations have a box size of $L_{box} = 95.5$ cMpc and assume a Planck Collaboration et al. (2016) cosmology. Our flagship simulation (THESAN-1) contains 2×2100^3 particles, reaching dark matter and gas resolutions of $m_{DM} = 3.12 \times 10^6$ M$_\odot$ and $m_{gas} = 5.82 \times 10^5$ M$_\odot$, respectively. This allows us to resolve atomic cooling haloes ($M_{halo} \sim 10^8 \, h^{-1}$ M$_\odot$). The forces are softened on a scale of 2.2 ckpc, while the smallest gas resolution elements reach sizes of approximately 10 pc. This run is accompanied by a number of medium-resolution runs, covering the same patch of the universe, but featuring slightly coarser resolution (i.e. 8 times lower mass resolution and 2 times lower spatial resolution). While the full description of these runs can be found in Paper I, we report in Table 1 the main features of the subset of runs employed in this Paper. In particular, the THESAN-2 run has the same physical model as the flagship one, while in THESAN-WC-2 we have adjusted the value of the stellar unresolved escape fraction to compensate the lower star-formation rate and keep the reionization history approximately identical to THESAN-1. In THESAN-LOW-2 and THESAN-HIGH-2 we force galaxies only above and below $M_{halo} = 10^{10}$ M$_\odot$ – respectively – to contribute to the EoR. In THESAN-SDAO-2, we replace the ΛCDM dark matter model with one featuring strong dark acoustic oscillations (e.g. Bohr et al. 2020). Finally, THESAN-TNG-2 employs the original IllustrisTNG model, which does not include full radiative transfer.

2.2 Synthetic spectra

The simulation outputs were post-processed to produce realistic synthetic spectra. We have developed a low memory method for exact ray-tracing through the native Voronoi unstructured mesh data as described in Paper II. We employ the Cosmic Lyα Transfer code (COLT; Smith et al. 2015, 2019) to extract the gas physical properties along

150 unique lines of sight (LOS) per snapshot. Each LOS has a length of 100 cMpc, originates from a randomly-chosen position in the simulation box, and propagates along a random direction employing periodic boundary conditions. These LOS are then processed to obtain the normalized transmitted flux for different transitions. We note here that we employ the full Voigt-Hjerting line profile (Hjerting 1938, using the approximation presented in Harris 1948 and Tepper-García 2006) including the effects of gas temperature and peculiar velocities and produce spectra with spectral resolution of $\Delta v = 1$ km s^{-1}. When comparing to a specific observational dataset, we additionally forward model these 'ideal' spectra by including noise (we make the assumption that the noise level is independent of the incoming flux, an approximation that is valid in the low-photons-count regime of the high-z Lyα forest, see Eilers et al. 2018), convolving the spectrum with a Gaussian kernel, and finally re-binning the spectrum to the appropriate spectral resolution. The parameters used in the procedure vary as we match them to those of the dataset employed in the comparison.

In order to characterise the high-z Lyα transmission, we follow the same procedure used in Garaldi et al. (2019b, itself adapted from Gnedin et al. 2017). In short, we identify regions of Lyα transmission (spikes) and characterise them by a height (h_s, corresponding to the maximum normalized transmitted flux), and a width (w_s, defined as the simply-connected set of pixels that have $f \geq \alpha h_s$, where α is a tuneable parameter that we set to $\alpha = 0.5$ in this work). To complement this analysis we quantify the distribution of low-flux regions using the dark gaps (dg) statistics (e.g. Paschos & Norman 2005; Fan et al. 2006; Gallerani et al. 2006, 2008), defined as contiguous regions where the (normalised) transmitted flux is below some threshold. Once the latter is fixed, dark gaps are characterised uniquely by their length (L_{dg}).

We stress here that, unlike many other numerical studies of the EoR (e.g. Becker et al. 2015, Bosman et al. 2018, and Eilers et al. 2019 using a uniform UVBG, Davies & Furlanetto 2016, Kulkarni et al. 2017 and D'Aloisio et al. 2018 using semi-analytical frameworks with multiple tunable parameters, and Kulkarni et al. 2019b, Keating et al.; Keating et al. 2018; 2020 using post-processing RT simulations) we do not tune our simulations to explicitly match Lyα or post-reionization IGM properties. In fact, the only tunable parameter of the THESAN runs, i.e. the stellar escape fraction, is set by requiring a late reionization history (see Paper I). Therefore, our predictions are directly and unequivocally linked to the physical modelling (including – as it is virtually always the case – free parameters tuned against some observations) employed in the simulations, that can consequently be tested in a more stringent way. We note here that, while tuning the stellar escape fraction to obtain a 'late' reionization history surely affects the post-reionization IGM, such a connection is not trivial, in particular because THESAN employs a much more sophisticated star formation and feedback model with respect to the original works suggesting that cosmic reionization is completed at $z \lesssim 5.5$ (i.e. Kulkarni et al. 2019b; Keating et al. 2018; Keating et al. 2020).

With the exception of hydrogen and helium, whose ionization states are self-consistently tracked in the simulations, we compute the ionization state of other elements using CLOUDY version 17.02 (Ferland et al. 2017). Following previous works (e.g. Bird et al. 2015; Nelson et al. 2018b) we employ the code in the single-zone mode (i.e. assuming that the gas density and temperature are constant across a single resolution element of the simulation) and iterate to equilibrium, including both photo- and collisional ionization. We include the radiation background of Faucher-Giguère et al. (2009, including its 2011 update) and account for gas self-shielding using
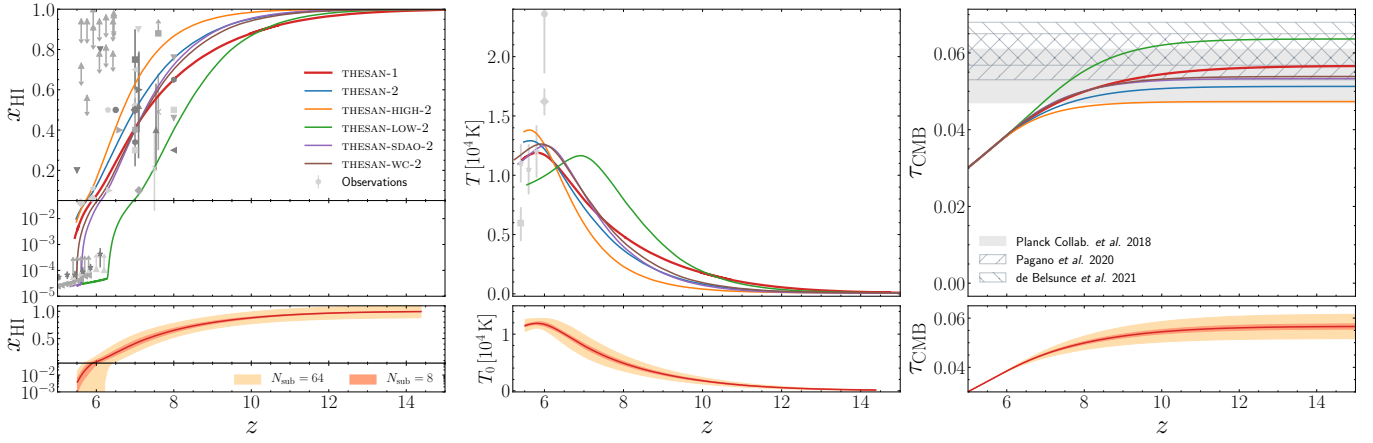
Figure 3. Global properties of the IGM in the THESAN simulations, compared to a collection of observations from the literature. From left to right: volume-averaged hydrogen neutral fraction evolution, volume-averaged temperature at mean density evolution, and CMB Thomson optical depth as a function of maximum integration redshift. The bottom panels show the variance around the THESAN-1 curve, computed by subdividing the original box into 8 (inner shaded region) and 64 (outer shaded region) sub-boxes. Overall, most THESAN simulations agree well with available data. Constraints on the neutral fraction are derived from the evolution of Lyα emitters and the number density of Lyman break galaxies (Ota et al. 2008 – light grey circles, Ono et al. 2012 – dark grey squares, Jensen et al. 2013 – dark grey circles, Pentericci et al. 2014 – light grey hexagons, Tilvi et al. 2014 – dark grey left-pointing triangles, Choudhury et al. 2015 – light grey squares, Mesinger et al. 2014 – grey circles), from the dark pixel statistics (McGreer et al. 2011 – dark grey down-pointing triangles, McGreer et al. 2015 – light grey diamonds, Lu et al. 2020 – grey up-pointing triangles), from GRB afterglow (Totani et al. 2006 – light grey pentagons, Chornock et al. 2013 – light grey stars, from the Doppler widths of Lyα absorption lines in the quasar near zones (Bolton et al. 2012 – light grey down-pointing triangles), from the quasar damping wing (Mortlock et al. 2011 – grey diamonds, Schroeder et al. 2013 – light grey right-pointing triangles, Greig et al. 2017 – dark grey right-pointing triangles, Greig et al. 2019 – light grey crosses, Wang et al. 2020 – light grey left-pointing triangles), from CMB modeling (Robertson et al. 2013 – dark grey diamonds), from the Gunn-Peterson optical depth (Fan et al. 2006 – dark grey stars, Davies et al. 2018 – dark grey up-pointing triangles, Yang et al. 2020 – light grey up-pointing triangles, Bosman et al. 2021 – grey left-pointing triangles), from the angular correlation function of Lyα emitters (Sobacchi & Mesinger 2015 – dark grey crosses), from the rest-frame UV continuum of galaxies (Schenker et al. 2014 – dark grey hexagons), from the detection of Lyα emission in Lyman break galaxies (Mason et al. 2018 – grey stars, Mason et al. 2019 – grey down-pointing triangles, Hoag et al. 2019 – grey squares, Jung et al. 2020 – dark-grey crosses), from a combination of Lyα luminosity function, clustering and line profile (Ouchi et al. 2010 – grey right-pointing triangles) and from the Lyα visibility (Dijkstra et al. 2011 – dark grey pentagons). Constraints on T_0 are derived from the Doppler widths of Lyα absorption lines in the quasar near zones (Bolton et al. 2010 – light-grey circles, Bolton et al. 2012 – light grey diamonds, from the Lyα forest power spectrum (Walther et al. 2019 – light-grey squares), and from the Doppler widths of Lyα transmission regions (Gaikwad et al. 2020 – light grey stars).

the prescription from Rahmati et al. (2013). We decided to proceed in this way, despite having self-consistent RT in the simulation, because of the broad frequency bins tracked by the code and relatively-narrow frequency range covered in comparison with the one required by CLOUDY. We do not include induced processes (Wiersma et al. 2009) and use solar abundances from Grevesse et al. (2010). We compute the abundances of metal ionization states as a function of hydrogen number density n_H, gas temperature T, and redshift. In particular, we sample the following ranges: $-7.0 < \log(n_H) < 4.0$, $0 < \log(T) < 8.5$ and $5.5 < z < 10$ on a Cartesian grid. We note that the metallicity dependence is minimal since we are just computing the *relative* abundance of each ionization state compared to the the neutral one for a given species. This procedure provides us the abundance $x_{i,j}$ of the j-th ion of the species i on a 3-dimensional Cartesian grid, which we then interpolate (via spline interpolation) to each combination of (n_H, T, z) extracted from the simulation. Therefore, the number density of a given ion species can be obtained simply as $n_{i,j} = n_i x_{i,j}$, where n_i is the number density of the metal species i, which is self-consistently tracked in THESAN for 9 metal species (see Sec. 2.1). This accurate forward modelling of the simulations allows us to faithfully study the properties of the high-z IGM, as presented in the following sections.

2.3 Comparison with other approaches

Before presenting the results concerning the IGM–galaxy connection, we take a moment to compare our methodology to a number of different works focusing on the EoR. We refer the reader to Paper I for a more thorough discussion and to Vogelsberger et al. (2020a) for an overview of numerical simulations of galaxy formation. Over the years, a multitude of approaches have been developed. Many of them (e.g. Keating et al. 2018; Keating et al. 2020; Kulkarni et al. 2017; Hassan et al. 2021) rely on completely de-coupling the radiation transport from the hydrodynamics and galaxy formation, treating the former in post-processing. While this significantly reduces their computational cost, and hence enables the exploration of larger volumes of the Universe, it is then not possible to capture the radiation feedback on galaxies and on the gas in the CGM and IGM. The latter is crucial to properly characterise the connection and interplay between galactic and cosmological scales, which is one of the primary goals of THESAN and the main focus of this paper.

In recent times, semi-analytical models like RSAGE (Croton et al. 2016; Hutter 2018; Seiler et al. 2019) and MERAXES (Mutch et al. 2016) have attempted to include radiation feedback without the computational burden of full radiation-hydrodynamical simulations. However, this comes at the price of an approximate calculation of both the galaxy properties and the radiation field.

Finally, fully-coupled radiation-hydrodynamical simulations of cosmic reionization have become available in recent times, often fo-

cusing on different selected aspects of the high-z Universe, a choice forced by the extreme computational cost of these calculations. For instance, SPHINX (Rosdahl et al. 2018) targets small scales, striving to resolve all the relevant photon sinks (hence removing the need for a stellar escape fraction) at the cost of relatively small box sizes. This is ideal for studying the internal properties of reionization sources, but the missing large scale coverage compromises the connection to cosmic reionization in a representative manner. The opposite approach is taken by the different iterations of the Cosmic Dawn project (Ocvirk et al. 2016, 2020), which simulates a significant volume of the universe but with coarse resolution (mainly as a consequence of the Cartesian grid employed). These calculations are well-suited to investigate IGM properties on cosmological scales (as e.g. the 21cm radiation topology), but lack the resolution needed to produce realistic galaxies and, hence, meaningfully study the IGM – galaxy connection. Finally, the CROC (Gnedin 2014) simulation suite is the closest one in terms of box size and resolution to THESAN. As we show in Paper I, however, high-mass galaxies in CROC deviate from the stellar-to-halo-mass relation inferred by observations and other cosmological simulations. As we will argue in Sec. 5, this may be the culprit behind the inability of these simulations to reproduce the observed modulation of Lyα flux with galaxy distance (Garaldi et al. 2019b), hence casting doubt on whether the IGM – galaxy correlation can be predicted faithfully. A further difference is that CROC is calibrated at high redshift, while THESAN is mostly calibrated at low redshift.

3 PROPERTIES OF THE IGM

We begin our analysis of the high-redshift IGM in the THESAN simulations focusing on its global properties. In order to provide an overview of the THESAN runs employed, we show in Fig. 1 the reionization redshift ($z_{\rm reion}$) in a slice through the different simulation boxes (whose name is reported in the top left corner of each panel). The largest panel shows our main fiducial run (THESAN-1). In order to compute $z_{\rm reion}$ we employ the so-called Cartesian outputs, which save at high time cadence a subset of gas properties on a Cartesian grid (see Paper I for details). For each pixel, we determine the local reionization history, which typically shows multiple periods of full reionization ($x_{\rm HII} \geq 0.99$) interspaced by periods of complete neutrality ($x_{\rm HII} \leq 0.01$). We therefore define $z_{\rm reion}$ as the minimum redshift such that $x_{\rm HII}(\tilde{z}) \geq 0.99$ for each $\tilde{z} < z_{\rm reion}$ (i.e. the last time the threshold value $x_{\rm HI} = 0.99$ is crossed from below). From a simple visual inspection, it can be appreciated that the reionization in THESAN proceeds inside-out, starting from the largest structures, which reionize their surroundings at $z \gtrsim 10$ (green to blue colors). Only much later do ionizing photons reach the lower-density IGM (yellow to red colors). The patchy nature of reionization is also evident from the Figure. Finally, the densest structures (e.g. individual galaxies and filaments) remain neutral until the final redshift of the simulations because of their high density and – consequently – large recombination rates. In the Figure, they stand out as dark red structures embedded in blue patches. While counter-intuitive, there is broad consensus that individual galaxies are simultaneously the sources of photons and the only structures in the Universe to remain neutral well beyond the end of the EoR. This is made possible by the multi-phase structure of their ISM and CGM, which combines high densities that boost the (hydrogen) recombination rate, keeping the overall H II fraction low, and highly-ionised channels that allow the escape of ionizing radiation in their surroundings (Wise & Cen 2009). As such, we caution the reader that the details of the ioniza-

tion state of these structures may depend on the specific ISM model employed in the simulations.

Some features of the different physical models can be readily appreciated by comparing the panels in Fig. 1. For instance, the bottom left panel displays how the IGM in THESAN-LOW-2 is reionized significantly earlier than in the other runs, as highlighted by the overall more yellow colors. Moreover, comparing it with the THESAN-2 and THESAN-HIGH-2 simulations, differences in the sources of radiation become apparent. These runs exhibit a smoother $z_{\rm reion}$ in the IGM, as a consequence of the fewer, larger ionization sources driving large coherent ionization fronts. On the contrary, THESAN-LOW-2 has the most inhomogeneous IGM. Nevertheless, the location of the earliest reionized patches is the same in all runs, showing that even when photons are forced to escape only from small galaxies, their clustering in overdense regions can produce an effect roughly comparable to our fiducial model.

By construction, $z_{\rm reion}$ condenses the information on the evolution of gas into a single number at each location. As such, it does not provide a straightforward insight on the IGM condition at any given redshift. Therefore, we complement it with Fig. 2, where we show the evolution in THESAN-1 of the ionising photons density and flux along a slice (of width equal to 20% of the box size). In particular, each panel shows (at a different redshift, reported in the top right corner) the linear integral convolution (LIC Cabral & Leedom 1993) of the photon density (acting as a background texture) and the (average) photon flux along the slice plane (providing the vector field). In order to better visualise the photon flux, we have added to the texture a small Gaussian noise (with zero mean). The color scale is the same in all panels. Hence, the varying hue of the picture is simply a reflection of the rapid evolution of the ionising background towards the end of the EoR. The Figure shows how galaxies and galaxy groups act as powerful sources of ionising photons, that travel approximately linearly until they are absorbed in a radiation sink.

We provide a more quantitative view of the global properties of the IGM in the top panels of Fig. 3, where we show as a function of redshift the volume-averaged hydrogen neutral fraction (left panel) and temperature at mean density (T_0, central panel). These panels reproduce and expands on figures 4, 5 and 6 of Paper I, and are included in order to provide a self-contained discussion. In the right panel, we show the CMB optical depth ($\tau_{\rm CMB}$) due to the integrated Thomson scattering of its photons off of free electrons – mainly produced by reionization – as a function of the uppermost redshift considered in its calculation. Below the final redshift $z_{\rm fin} = 5.5$ of the simulation we assume for this calculation that (i) hydrogen is fully ionized, (ii) helium is singly ionized at $3 < z \leq 5.5$ and doubly ionized at $z \leq 3$. In all cases, we report a collection of constraints from the literature. The reionization history of the THESAN runs is consistent with the bulk of observations, derived from the evolution of Lyα emitters and the number density of Lyman break galaxies (Ota et al. 2008; Ono et al. 2012; Pentericci et al. 2014; Choudhury et al. 2015; Tilvi et al. 2014; Mesinger et al. 2014), from the dark pixel statistics (McGreer et al. 2011, 2015; Lu et al. 2020), from GRB afterglow (Totani et al. 2006; Chornock et al. 2013), from the Doppler widths of Lyα absorption lines in the quasar near zones (Bolton et al. 2012), from the quasar damping wing (Mortlock et al. 2011; Schroeder et al. 2013; Greig et al. 2017; Greig et al. 2019; Wang et al. 2020), from CMB modeling (Robertson et al. 2013), from the Gunn-Peterson optical depth (Fan et al. 2006; Davies et al. 2018; Yang et al. 2020; Bosman et al. 2021), from the angular correlation function of Lyα emitters (Sobacchi & Mesinger 2015), from the rest-frame UV continuum of galaxies (Schenker et al. 2014), from the detection of Lyα emission in Lyman break galaxies (Mason et al. 2018; Mason et al. 2019;
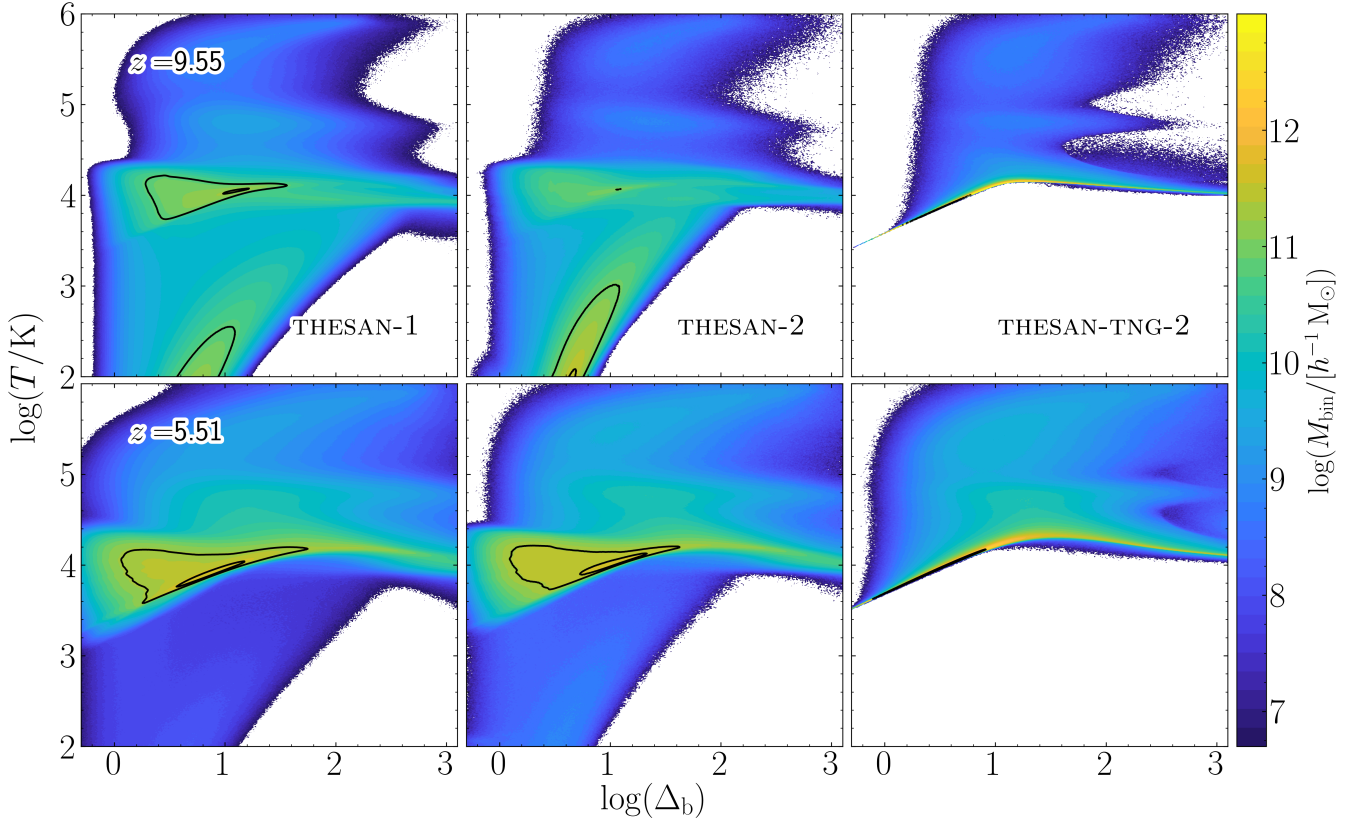
Figure 4. Temperature-baryon overdensity phase space of the gas in the THESAN-1 (left), THESAN-2 (center) and THESAN-TNG-2 (right) runs during the initial ($x_{\rm HI} \gtrsim 0.9$, $z \sim 10$, top) and final ($x_{\rm HI} \lesssim 0.01$, $z \sim 5.5$, bottom) phases of reionization. The color-coding scales logarithmically with the mass in each pixel. Note that, in order to highlight the differences in the intergalactic medium, we plot only the most relevant part of the phase space. It can clearly be seen that the inclusion of self-consistent radiation transport and non-equilibrium thermo-chemistry greatly alter the properties of the low-density warm IGM, providing a significantly more realistic description than through the use of an uniform UVBG.

Hoag et al. 2019; Jung et al. 2020), from a combination of Lyα luminosity function, clustering and line profile (Ouchi et al. 2010) and from the Lyα visibility (Dijkstra et al. 2011. This is by construction, as we explicitly tune the stellar (unresolved) escape fraction of the simulations to obtain a realistic reionization history. We note here that all runs in the THESAN suite except THESAN-LOW-2 follow a late reionization model, necessary to account for the $z \sim 5.5$ opacity of the Lyα forest (see Kulkarni et al. 2019b; Keating et al. 2020, and the discussion in Section 4). We refer the reader to Paper I for a more in-depth discussion. The (very scarce at $z \gtrsim 5.5$) constraints on T_0 are derived from the Doppler widths of Lyα absorption lines in the quasar near zones (Bolton et al. 2010, 2012), from the Lyα forest power spectrum (Walther et al. 2019), and from the Doppler widths of Lyα transmission regions (Gaikwad et al. 2020). These constraints are mutually inconsistent (as they would require an unphysically rapid cooling of the IGM). The THESAN runs all lie between the observed values, with a mild dependence of the maximum T_0 on the duration of the reionization process (see D'Aloisio et al. 2019, and Paper I for a more detailed discussion). Because of its earlier reionization, the IGM temperature peaks earlier in THESAN-LOW-2 compared to the other runs. Finally, the CMB optical depth predicted by the THESAN runs agrees very well with the estimation from Planck Collaboration et al. (2020). Only THESAN-LOW-2 features too large of a $\tau_{\rm CMB}$, again as a consequence of its earlier reionization. However, recent re-analysis of the Planck Collaboration et al. (2020) data by Pagano

et al. (2020) and de Belsunce et al. (2021) found larger values of $\tau_{\rm CMB}$. While THESAN-1 is consistent with (although on the low end of) these two estimates, many of the other THESAN runs are not.

In the bottom panels of Fig. 3 we investigate the effect of the box size on the global IGM properties, providing indications on the relevance of sampling variance in this study. For the sake of clarity, we only investigate this in the THESAN-1 run. In particular, we have divided the simulation box in two sets of $N_{\rm sub} = 8$ and $N_{\rm sub} = 64$ sub-boxes. As references, an individual sub-box in the latter set has a linear size twice as large as the fiducial SPHINX simulation (Rosdahl et al. 2018). Within each set, the union of the sub-boxes covers the entire volume. We compute the redshift-dependent values of $x_{\rm HI}$ and T within each sub-box, and plot their scatter around the full-box average as a shaded region in the bottom panels of Fig. 3. It can be seen that, while the scatter obtained using 64 sub-boxes is significant, it is almost-entirely suppressed when $N_{\rm sub} = 8$. While conclusive statement about the impact of sample variance cannot be made without running a larger equal-resolution box (which is currently unfeasible because of numerical limitations), these results are encouraging. In fact, each sub-box of the $N_{\rm sub} = 8$ set already shows a global evolution hardly distinguishable from the full-box one, indicating that the role of the box volume is marginal for volumes $V \gtrsim (L_{\rm box}/2)^3$. Therefore, we expect sample variance to play an equally-marginal role in the result presented in this work.

Reassured by this finding, we continue our investigation of the

IGM properties by showing in Figure 4 the joint distribution (with colors reflecting the mass in each pixel) of gas temperature and baryon overdensity. The latter is defined $\Delta_b \equiv \rho_b/\bar{\rho}_b$, where ρ_b is the baryon density (that in the IGM coincides with the gas density) and $\bar{\rho}_b$ is its average value. The top row of the Figure displays the gas state at $z = 9.5$, when the reionization process is still in its early stages, while the bottom row shows the gas state reached at the tail-end of the EoR ($z = 5.5$), where most of the gas in the simulation has already been fully ionized. The Figure also presents a comparison between THESAN-1 (our flagship simulation, left column), THESAN-2 (employing the same physical model at 8× lower mass resolution, central column), and THESAN-TNG-2 (employing the original IllustrisTNG model to evolve the same initial conditions as THESAN-2, right column). Finally, each panel shows black contours encompassing the central 68%, 95% and 99.5% of the data. Before presenting our interpretation of these results, we note here that the range of baryonic overdensities and temperatures reached in our simulations is much larger than shown in the panels. In particular, in the THESAN simulations temperatures down to 5 K and up to $\sim 10^8$ K are found, as well as baryonic overdensities up to $\Delta_b \sim 10^8$. However, we choose to focus here on the most relevant region of phase space for the IGM – the main topic of this paper – during and after reionization. In addition, at large densities, the gas follows an effective equation of state (Springel & Hernquist 2003), which suppresses most of the excursion in temperature.

A first striking difference can be appreciated comparing the right column with the others. The original IllustrisTNG model does not self-consistently track radiation, as THESAN does, but rather employs a spatially-uniform UV background (Faucher-Giguère et al. 2009). This approach results in a very tight temperature-density relation with little redshift evolution. Conversely, a more accurate treatment of radiation (including photo-heating and non-equilibrium thermo-chemistry) allows us to fully capture the inhomogeneous process of reionization, which can be seen in the left and central columns as a large distribution of temperatures associated with a single density value. This is the result of the overlap of individual regions ionized at different times. Each of them establishes its own temperature-density relation as a consequence of photo-heating, which then evolves because of gas cooling (via adiabatic expansion and Compton cooling; e.g. Trac et al. 2008; Furlanetto & Oh 2009). We note here that, while at $\log \Delta \lesssim 1$ the RT is responsible for the differences, the right-hand part of the diagram is ruled by hydrodynamical and galactic effects. Therefore, the junction of these two regimes can be faithfully captured only by combining self-consistent RT with an accurate galaxy-formation model, as in THESAN. For instance, employing post-processing RT, even on finely-spaced outputs from hydrodynamical simulations, results in a mismatch in the $\log \Delta \gtrsim 1$ region of the phase space (Keating et al. 2018). Employing a spatially-uniform UV background (as routinely done in large hydrodynamical simulations and as done in THESAN-TNG-2) entirely suppresses the low-temperature gas found during reionization when RT is self-consistently implemented. (We note here that, additionally, the THESAN simulations employ a non-equilibrium solver for the thermal state of the gas, unlike the IllustrisTNG simulations.) The high-temperature portion of this diagram (corresponding to $T \gtrsim 10^{4.5}$ K) is remarkably similar in all the runs presented in the Figure, since gas is brought to such a thermal state by energetic feedback processes originating from galactic physics, which is unchanged in the three runs. Finally, we note that there are small differences in the IGM structure at $z = 9.55$ between THESAN-1 and THESAN-2. These are a consequence of the slightly delayed reionization history in the latter (see Fig. 3), which results in a difference in the phase space when compared at the same redshift.
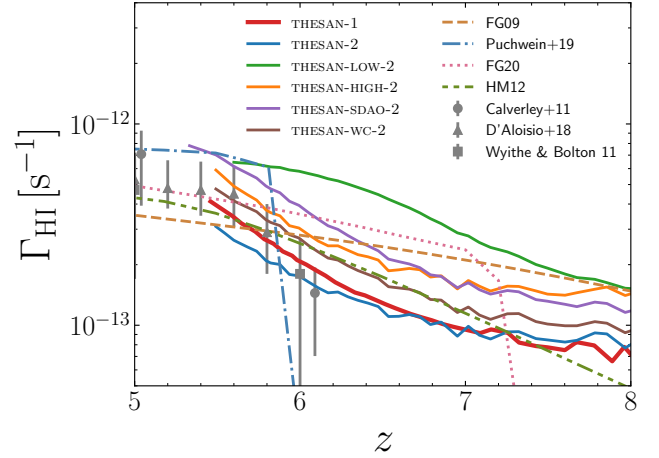
Figure 5. Photo-ionization rate for HI computed in ionized regions (i.e. where $x_{HII} \geq 0.5$). Solid lines show the results extracted from different THESAN runs, the dashed, dot-dashed, dotted and double-dashed lines report the predictions from the UVBG models of Faucher-Giguère et al. (2009, including its update in December 2011), Puchwein et al. (2019), Faucher-Giguère (2020), and Haardt & Madau (2012), respectively. Symbols show measurements from Calverley et al. (2011); D'Aloisio et al. (2018); Wyithe & Bolton (2011).
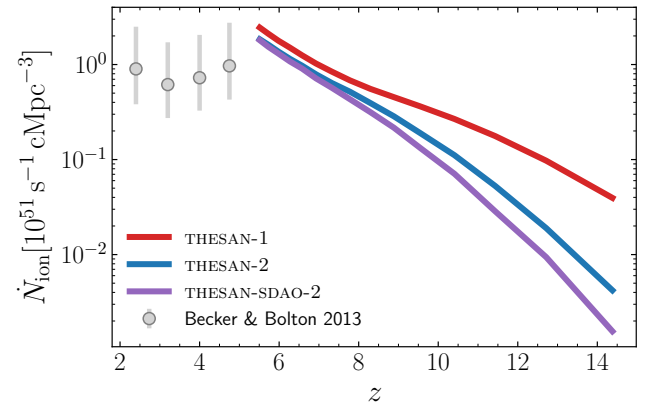
Figure 6. Evolution of the photon emission rate from stars in the THESAN-1, THESAN-2 and THESAN-SDAO-2 runs, compared to observations of Becker & Bolton (2013). Since this quantity depends only on the stellar population in the simulations, the THESAN runs not shown here overlap almost perfectly with THESAN-2.

The density and temperature of the IGM crucially depends on the position- and frequency-dependent radiation field. We exploit here the self-consistent RT to investigate the accuracy of available UVBG models. We begin by showing in Figure 5 the H I photo-ionization rate Γ_{HI}, computed in ionized regions (defined as gas cells with neutral hydrogen fraction $x_{HI} < 0.5$) as a function of redshift. The solid lines in the Figure show the outcome from a selection of the THESAN simulations, while symbols report values observed by Calverley et al. (2011), D'Aloisio et al. (2018), and Wyithe & Bolton (2011). Almost all THESAN runs are compatible with available data. However, it appears clear that the simulated curves have a shallower evolution than observed. In fact, while data seem to rise by a factor ~ 10 between $5.5 \leq z \leq 6$ and then flatten, in simulations Γ_{HI}
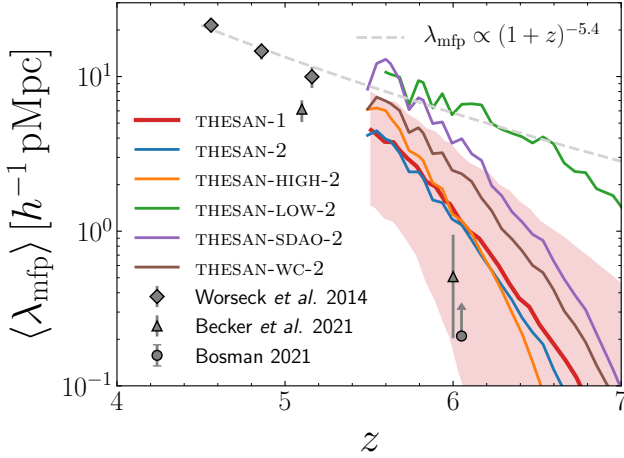
Figure 7. Mean free path of ionizing photons as a function of redshift, compared to observations from Worseck et al. (2014), Becker et al. (2021) and Bosman (2021, displaced by $\Delta z = 0.05$ for visual clarity). The dashed line shows the fit to $z < 5.5$ data reported in Worseck et al. (2014). THESAN-1 reproduces the recently observed rapid growth of the mean free path, as a consequence of its late reionization history. Medium resolution simulations show a similar evolution, except for THESAN-LOW-2, which features a much earlier reionization history, close to the classical picture.

steadily grows since $z \sim 8$ and shows no sign of stagnated growth by $z = 5.5$.

Interestingly, the only THESAN run largely not matching the available data is THESAN-LOW-2. This is a consequence of the earlier completion of the reionization process in this case. We can try to approximately account for the different reionization history of THESAN-LOW-2 by replacing the original curve $\Gamma(z)$ with $\Gamma(\tilde{z})$, where $\tilde{z}$ is the redshift that matches the reionization history of THESAN-LOW-2 with the one of THESAN-1, i.e. such that $x_{\mathrm{HI}}^{\mathrm{THESAN-LOW-2}}(\tilde{z}) = x_{\mathrm{HI}}^{\mathrm{THESAN-1}}(z)$. This corresponds approximately to a delay of $\Delta z \sim 0.8$ and brings THESAN-LOW-2 in better agreement with (although still partially higher than) THESAN-HIGH-2 and THESAN-SDAO-2. These three curves lie consistently above the ones from THESAN-1 and THESAN-2, and partially the one corresponding to THESAN-WC-2, and the reason can be found in the different number of stellar sources effectively contributing to the EoR. In fact, while in the fiducial physical model every star particle injects photons in the local environment, we suppress the photon escape from low-mass galaxies in THESAN-LOW-2 and from high-mass ones in THESAN-HIGH-2. Similarly, the THESAN-SDAO-2 run has modified initial conditions that suppress small scale power, reducing the number of small galaxies. Hence, for their reionization history to be consistent with available data, the stellar escape fraction f_{esc} needs to be higher in these models, in order to counter-balance the reduction of ionizing sources, effectively producing a larger photon density (and hence Γ_{HI}) in the ionized regions.

Another significant conclusion that can be extracted from the analysis of Figure 5 is that THESAN-1 and THESAN-2 show very similar values of Γ_{HI} throughout most of the simulation, indicating good convergence of the two models. Only below $z \lesssim 6$ does THESAN-2 start to lag behind THESAN-1, and this is reflected in its slightly-delayed reionization history. Additionally, when compared to values inferred from observations, the fiducial physical model fits the data best among the THESAN runs, as the other are generally higher than the data at $z \gtrsim 5.8$. Nevertheless, the observed Γ_{HI} shows a steep increase (by almost an order of magnitude) between $z = 6.1$ and

$z = 5.5$, and a subsequent flattening at lower redshift. Conversely, the evolution in all the THESAN runs is significantly slower and its slope does not become shallower down to $z \sim 5.5$. This may be indicative of a reionization history that is too extended in our simulation. There are, however, other explanations. For instance, the number of haloes above $M_{\mathrm{thr}} = 10^{10}\,\mathrm{M}_\odot$ increases with time. Therefore, the number of haloes contributing to reionization in THESAN-LOW-2 decreases at lower redshift, explaining the flattening in this model. Finally, we compare the simulated evolution of Γ_{HI} with a selection of UVBG models available in the literature. The dashed, dot-dashed, dotted and double-dashed lines in the Figure report the predictions from Faucher-Giguère et al. (2009, FG09, including its update in December 2011), Puchwein et al. (2019), Faucher-Giguère (2020, FG20), and Haardt & Madau (2012, HM12), respectively. Surprisingly, the model that most closely follows the results from the THESAN-1 and THESAN-2 runs is the one from HM12, while FG09 more closely resembles the results from THESAN-HIGH-2 and THESAN-SDAO-2.

Before moving on to the characterisation of the mean free path of ionizing photons, we show in Fig. 6 the evolution of the volume-averaged ionizing photon emission rate in THESAN-1, THESAN-2 and THESAN-SDAO-2, compared to available measurements from Becker & Bolton (2013). Note that other medium resolution runs are virtually identical to THESAN-2 in this regard, since their star-formation history is only very mildly affected by the different radiation fields. At very high redshift, THESAN-1 produces a significantly larger number of ionizing photons than THESAN-2, thanks to its higher resolution able to resolve smaller objects (see also figure 12 in Paper I). At $z \lesssim 8$, these two runs show a similar rate of injection of ionizing photons. Similarly, the suppression of some small-scale modes in the THESAN-SDAO-2 initial conditions reduces the number of ionizing photons emitted at $z \gtrsim 8$ with respect to THESAN-2. By redshift $z \sim 8$, however, $\dot{N}_{\mathrm{ion}}$ is similar in all runs. If extrapolated to $z \leq 5$, the value would be slightly larger than observed by Becker & Bolton (2013), showing that an inversion in this trend is needed to match the observed constraints. In Kulkarni et al. (2019a) and Keating et al. (2020) such a sharp decrease in $\dot{N}_{\mathrm{ion}}$ had to be assumed in order to match the mean Lyα flux evolution, although these same authors claim it is very challenging to find a plausible physical interpretation. Ocvirk et al. (2021) found that a strict upper limit on the temperature of the gas allowed to form stars can reproduce such a drop.[2] Our simulations adopt a similar approach, preventing gas with temperatures above the one predicted by its effective equation of state (e.g. as a consequence of heat injection from feedback processes) from forming stars. Despite this, we do not see any sign of decrease in $\dot{N}_{\mathrm{ion}}$ at $z \lesssim 6$. The reason for this difference may lie in the fact that feedback processes typically heat the gas to temperatures that are orders of magnitude higher than the one employed by Ocvirk et al. (2021) as an upper temperature threshold.

3.1 Mean free path evolution

Very recent measurements of the mean free path of ionizing photons (λ_{mfp}) from the attenuation of high-z quasar spectra point to a sharp drop in this quantity between $z \sim 5$ and $z \sim 6$ (Becker et al.

---

[2] We note here that the authors conclude that the drop in $\dot{N}$ is a consequence of the photo-ionization feedback from percolating reionized bubbles. However, the different distribution of stars produced by the two star-formation prescriptions likely creates significant differences in the stellar and supernova feedback on the gas, which may produce effects on star formation at least as strong as photo-ionization feedback.
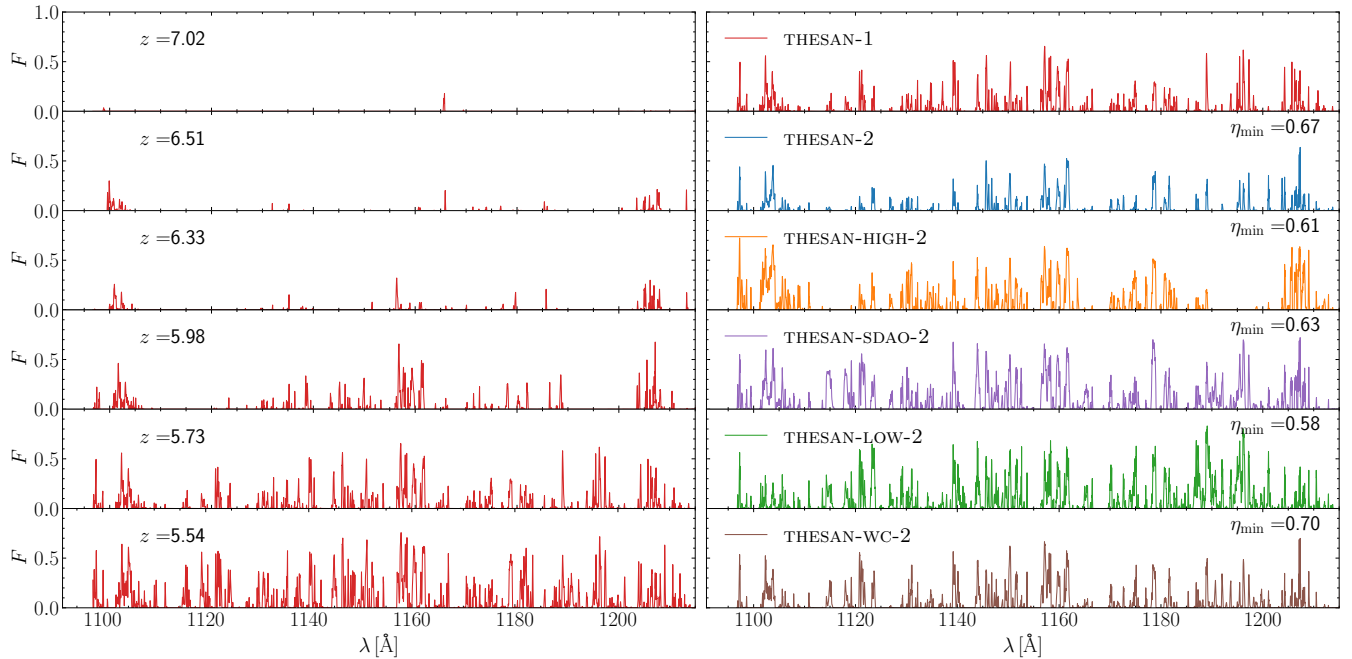
Figure 8. Example of forward modeling synthetic spectra in the THESAN suite. The left column shows the evolution with redshift of the same line of sight in the THESAN-1 flagship run. The right column displays the different Lyα spectra obtained at the same redshift ($z = 5.73$) in a selection of THESAN runs. The vertical axis shows the normalized transmitted flux while the horizontal one reports the rest-frame wavelength. In the top right corner of the right-hand-side panels, we report the overlapping index η_{min} of the sightlines in the run shown and in THESAN-1.

2021; Bosman 2021). This newly-observed λ_{mfp} lies significantly below the value at the same redshift predicted by extrapolating measurements at $4 \lesssim z \lesssim 5$ (Worseck et al. 2014). We compare these new measurements with the predictions from the THESAN suite in Figure 7, where we also show the distribution of the central 68% of ionizing photons free paths computed from the simulations in THESAN-1 (other THESAN runs have similar scatter). We stress that the simulations were *completed* before such measurements were made public. In order to compute λ_{mfp} from the simulated IGM, we follow Rahmati & Schaye (2018) and extract random (both in origin and direction) skewers through the simulated volume and determine the mean free path, λ_{mfp}, from the following definition:

$$\int_0^{\lambda_{\mathrm{mfp}}} \mathrm{d}s\, \sigma_{912}\, n_{\mathrm{HI}} = 1 \,, \tag{1}$$

where σ_{912} is the cross-section for photons with wavelength $\lambda = 912\,\text{Å}$ and $\mathrm{d}s$ is the line-of-sight distance. The predicted evolution of λ_{mfp} in our fiducial model is in excellent agreement with the measurements from Becker et al. (2021) and the lower limit from Bosman (2021) at $z \sim 6$ and appear to approach the available data at $z \sim 5$. The same holds true for the THESAN-2 and THESAN-HIGH-2 runs, which are in very good agreement with the results of our flagship run. The agreement between THESAN-1 and THESAN-2 appears here somewhat fortuitous, since these two runs have both a different reionization history and different resolution (i.e. resolve small-scale structures to a different degree). This can be seen comparing them with THESAN-WC-2, which has a reionization history very close to the one of THESAN-1 and shows – on average – a larger mean free path. Additional evidence in support of this interpretation comes from the inspection of the evolution of λ_{mfp} with the ionized fraction (instead of redshift, removing the dependence on the reionization history, not shown for the sake of brevity), which appears indistinguishable

in THESAN-2 and THESAN-WC-2, but offset toward lower values in THESAN-1. This also demonstrates that the THESAN simulations are not completely converged in their predictions of λ_{mfp}. It appears likely that, in order to match the median λ_{mfp} in THESAN-1 with the central value obtained by Becker et al. (2021), a slightly delayed reionization history is required.

Differently from the other runs, THESAN-LOW-2 shows a much larger mean free path at all redshifts, as a consequence of its significantly earlier reionization history. Note that once reionization is completed in this model, its λ_{mfp} settles on the extrapolation of the Worseck et al. (2014) observations, suggesting that in THESAN-LOW-2 the post-reionization mean free path is compatible with the observed one. Most of the difference can be ascribed to the early reionization history of this run. However, even after accounting for such effects (see above), there are residual differences among THESAN-2, THESAN-LOW-2 and THESAN-HIGH-2. In particular, the latter has a larger mean free path than our fiducial physical model, while the photon free path in THESAN-LOW-2 is systematically lower than in THESAN-2 (similarly to the recent results of Cain et al. 2021). Finally, THESAN-SDAO-2 exhibits values of the mean free path that are a few times larger than the observed ones at $z \sim 6$, and slightly larger than THESAN-WC-2, again as a consequence of the slightly earlier completion of reionization (we have checked this is the correct interpretation by inspecting $\lambda_{\mathrm{mfp}}(x_{\mathrm{HI}})$). It is interesting to note that, even with the strong suppression of certain (low) mass haloes in the sDAO initial conditions compared to the CDM ones, the mean free path is not significantly affected (once the slightly different reionization histories are accounted for).

Recently, Cain et al. (2021) employed a coarse-grained RT simulation coupled with higher-resolution small-volume runs to approximately resolve small-scale structures in large volumes, and concluded

that: (i) the rapid evolution in the mean free path observed by Becker et al. (2021) favours a reionization driven by small, high-escape-fraction galaxies, and (ii) additional sinks are needed beyond those predicted by high-resolution numerical simulations to match the central value obtained by Becker et al. (2021). The results presented here partially support the first conclusion, showing that a reionization epoch driven by the smallest (largest) galaxies produces a shorter (larger) mean free path (at fixed reionization history). However, the ionizing photon budget in THESAN-1 is dominated by galaxies residing in haloes with mass $M_{\rm halo} \geq 10^{10}\,{\rm M}_\odot$ at $z \leq 8$ (see Figure 12 in Paper I). Additionally, our fiducial model is capable of recovering the $\lambda_{\rm mfp}$ evolution in a way similar to Cain et al. (2021) in a self-consistent RHD simulation framework. We note here that THESAN-1 resolves haloes with mass $M_{\rm halo} \gtrsim 10^8\,{\rm M}_\odot$ (this value is 8 times larger for the lower-resolution runs). These are smaller haloes than the smaller sub-resolution sinks in the fiducial configuration of Cain et al. (2021). Additionally, our simulations can resolve sinks of significantly smaller mass if they are gas dominated, thanks to the better gas mass resolution with respect to that of the DM.

In Davies et al. (2021), the recent measurements of the mean free path at $z \sim 6$ were combined with the coeval dark pixel fraction to infer a cumulative output of ionizing photons per baryon $n_{\gamma/\rm b} = 6.1^{+11}_{-2.4}$, a number significantly larger than typically assumed. Moreover, these high values require an escape fraction $\gtrsim 20\%$ across the entire galaxy population. In THESAN-1, we find that $n_{\gamma/\rm b} = 6.31$ photons per baryon are needed to reionize the IGM, in agreement with the value quoted above. In order to compute this number, we consider the total ionizing output from stars and include an effective escape fraction of photons into the IGM, which includes both the unresolved escape from the birth cloud, modelled through the $f_{\rm esc}$ parameter, and the resolved absorption from the ISM and CGM of simulated galaxies, resulting in an effective escape fraction of order $\sim 10\%$ for most of the simulation (see the right panel of Figure 17 in Paper I and the relative discussion therein).

Studying the IGM is very challenging, especially at high-z, because its low density renders its emission very dim, leaving absorption studies as the only available approach. The main absorption line studied in this context is Lyα, which we investigate next.

4 THE HIGH-REDSHIFT LYα FOREST

We now move to an in-depth analysis of one of the main observables of the EoR, the Lyα transition of neutral hydrogen. We have described in Section 2.2 the process for the production of synthetic absorption spectra that are employed here, and we show in Fig. 8 an example of the resulting spectra. In particular, the left column shows the evolution of the transmitted flux (as a function of rest-frame wavelength) along the same line of sight extracted from the THESAN-1 run. As expected, the number of transmission regions grows with redshift. In the right column we show the same LOS at $z = 5.73$ in the different runs employed in this Paper (and summarised in Table 1). While the position and characteristics of the transmission features are overall similar, small differences among the different runs exist, which warrant the further in-depth analysis that we carry out in this Section. Before proceeding further, we quantify these differences in the right column of Fig. 8 using the overlapping coefficient

$$\eta_{\rm min}(f, g) \equiv \int \min[f(x), g(x)]\, {\rm d}x \left/ \int {\rm d}x \right., \qquad (2)$$

where f and g are two arbitrary functions with unit integral and the integration runs over their entire (common) support. $\eta_{\rm min}$ is unity

in case of perfect overlap and vanishes in case at least one function is vanishing at each common support point. In our case, we fix f to be the (appropriately normalised) flux in the THESAN-1 run (top right panel) and g to be the (appropriately normalised) flux along the same sightlines in the different run shown. We report the value of this overlap coefficient in the top right corner of the right-column panels.

The simplest measure of the high-redshift Lyα transmission is the average transmitted (normalised) flux $\bar{F}$, shown in Figure 9. Symbols report values computed from observed high-z quasar spectra (Fan et al. 2006; Becker et al. 2015; Bosman et al. 2018; Eilers et al. 2018; Yang et al. 2020; Bosman et al. 2021), while solid lines show the values computed from the THESAN suite. For the sake of clarity, we show the central 68% of the data with a shaded region only for THESAN-1, as the other runs show a similar scatter. This simple plot already reveals interesting patterns. First, similarly to the cases discussed in Section 3, the runs employing our fiducial physical model (i.e. THESAN-1 and THESAN-2) are most consistent with data. Despite the extreme sightline-to-sightline variation in F and the consequent scatter, the various simulations show consistent differences that remain across all investigated redshifts. In particular, all but THESAN-LOW-2 show a virtually identical redshift evolution with different amplitudes. Taking THESAN-1 as a reference, we now explore such trends. Sightlines in THESAN-2 typically show a slightly lower transmitted flux, consistent with its somewhat delayed reionization history, which is a consequence of the smaller number of low-mass galaxies resolved in this simulation (see e.g. Fig. 6). Conversely, THESAN-HIGH-2 exhibits an enhancement in $\bar{F}$ with respect to THESAN-1. This can be readily explained by the fact that in this run ionizing photons are only emitted by large galaxies, while the stellar escape fraction is increased to complete reionization at approximately the same time as THESAN-1. Therefore, in this run ionized regions are typically larger and have higher ionized fractions, overall producing an enhancement of the transmitted flux. In other words, since we explicitly adjusted $f_{\rm esc}$ to maintain the volume-averaged ionized fraction approximately the same, the concentration of photons around a small number of bright sources produces a larger transmission when compared to the same IGM location in the THESAN-1 run. This can be seen in Figure 8, comparing the first and third panels from the top in the right column. The former shows slightly more transmission regions than the latter (notice e.g. the second group of transmission peaks from the right, which is present in THESAN-1 but absent in THESAN-HIGH-2) because of the larger number of ionized regions, but where both spectra have non-zero transmission, THESAN-HIGH-2 shows typically a larger flux (appreciable e.g. in the leftmost group of high-transmission regions). We will explore the feasibility of using the statistics of Lyα transmission regions to constrain the sources of reionization in a forthcoming work.

Figure 9 additionally shows that there exists an offset between THESAN-SDAO-2 and THESAN-1, similar to the one described above but larger in size. In the former run, in fact, the values of $\bar{F}$ are significantly larger than in the latter at all redshifts. This is also reflected in its earlier completion of reionization. An extreme case is represented by THESAN-LOW-2, which shows a much larger value of $\bar{F}$ that additionally evolves more slowly with redshift and remains almost constant below $z < 6$, in stark contrast with available observations. It is important to note here that, unlike in previous discussions, a naïve translation of the curve to lower redshifts to account for the earlier reionization in THESAN-LOW-2 does not bring this run into agreement with observations. In fact, in the latter, the value of $\bar{F}$ increases more than linearly with redshift, while THESAN-LOW-2 exhibits the opposite behaviour. The reason for this lies in the fact that the contri-
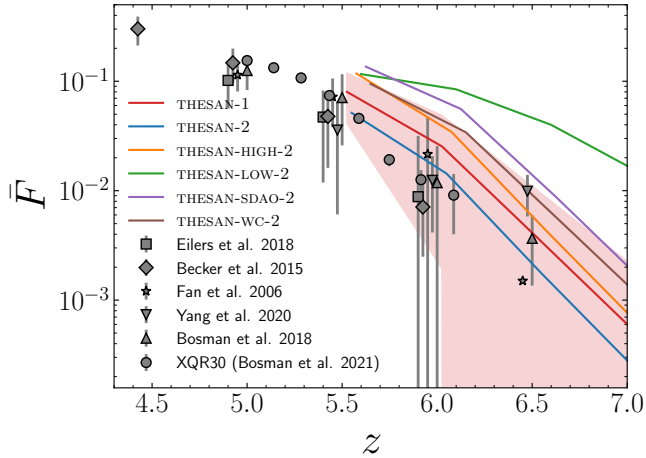
Figure 9. Mean transmitted flux in the Ly α forest as a function of redshift for a selection of THESAN runs, compared to the observations of Fan et al. (2006), Becker et al. (2015), Bosman et al. (2018, 2021), Eilers et al. (2018), and Yang et al. (2020). For clarity, we show the central 68% of the data for THESAN-1 only as a shaded region. The other simulations exhibit a similar scatter.

bution of small galaxies (i.e. with DM-halo mass $M_h \leq 10^{10}\,M_\odot$), the only ones injecting ionizing photons in the IGM in this model, decreases significantly towards the end of the EoR, ultimately stagnating reionization. This can quantitatively be seen by looking at the contribution to the global star formation-rate density from galaxies residing in haloes of different mass. This quantity is shown in Figure 12 of Paper I for e.g. the THESAN-2 run, whose galaxy and stellar populations are virtually identical to those in THESAN-LOW-2.

Finally, our flagship simulation THESAN-1, while consistent with available measurements of $\bar{F}$, appears to lie slightly higher than the latter at $z \lesssim 6$. However, the extreme difficulties linked to measuring the Ly α transition at such high redshifts, the patchy nature of reionization, as well as the limited availability of sightlines, render these measurements subject to large uncertainties. Hence, we believe that our fiducial model provides a realistic description of Ly α transmission during the reionization epoch.

4.1 Effective optical depth distribution

The trends described above are also reflected in the evolution of the effective optical depth $\tau_{\mathrm{eff}} \equiv -\ln\langle F \rangle$ (where $\langle F \rangle$ is the mean flux in spectral segments of length $50\,h^{-1}$ Mpc), as shown in Figure 10. Shaded regions display the central 95% of the distributions extracted from a selection of the THESAN simulations, while symbols and arrows report measurements and lower limits – respectively – from Fan et al. (2006), Becker et al. (2015), Bosman et al. (2018, 2021), Eilers et al. (2018), and Yang et al. (2020). With the exception of THESAN-LOW-2, which exhibits significantly lower optical depth with respect to the other runs at all redshift, the predictions for all other THESAN runs are very consistent with each other. Additionally, they match the observed τ_{eff} remarkably well at all redshifts. At the highest redshifts, most observations provide only lower limits, since the flux in the Ly α forest is below the detection threshold. This results in an apparently stalled evolution of τ_{eff}, which our simulations predict to simply be an artefact of the increasing difficulty of detecting the diminishing transmitted flux in observations.

With this caveat in mind, we now move to a more quantitative analysis of the cumulative distribution function (CDF) of effective optical depths at a fixed redshift in Figure 11. It shows the cumulative distribution of τ_{eff} in four different redshift bins, centred on $z = 5.45, 5.75, 6.05, 6.45$ with a width of $\Delta z = 0.3$. The choice of these redshift bins is motivated by the requirement of having a sufficient number of observed sightlines in each bin to meaningfully compare with the simulated ones. We report their average number across the different datasets in the top left corner of each panel, together with the redshift covered by the bin. For each dataset, we estimate the distribution of CDF via bootstrapping, creating 500 realizations of the sample and showing the central 68% of them. In addition, only in the case of THESAN-1, we force each bootstrapped sample to contain a number of sightlines equal to the average number in the observed dataset, in order to account for sample variance. We account for the finite sensitivity of the observations by assuming that the intrinsic flux equals $\max(\langle F \rangle, 2\sigma_{\langle F \rangle})$, where $\sigma_{\langle F \rangle}$ is the error on $\langle F \rangle$. A slightly different strategy is taken for the Bosman et al. (2018) and XQR30 (Bosman et al. 2021) datasets, in order to follow the approach of the original authors. In fact, they provide an 'optimistic' and a 'pessimistic' method for the handling of low-flux pixels. The former is identical to the procedure just described. The latter, instead, assumes the intrinsic flux is vanishing (i.e. $\tau_{\mathrm{eff}} \to \infty$). Finally, we do not attempt to model observational errors on the transmitted flux in this context, since they are different for each (and within each) dataset, preventing a fair comparison. However, Eilers et al. (2018) showed that – for Ly α – their effect is typically to slightly steepen the high-τ_{eff} end of the CDF at redshift $z \gtrsim 6$. This is a consequence of the fact that whenever the flux is below twice the associated error, the latter is assumed while computing the optical depth. Hence, including error degrades the sensitivity towards high optical depth, effectively placing an upper limit on the values of τ_{eff} that can be obtained. We have checked that in our case this effect is negligible. With the exception of the highest-redshift bin, the results from THESAN agree with the observed distribution of optical depth, although they lie on the lower τ_{eff} side of the observed values. This may indicate that the Universe is reionized even later than what is predicted by THESAN-1, or that our resolution is preventing us from resolving some of the densest structures, acting as sinks of photons.

We can test these effects by comparing the predicted CDF in THESAN-1, THESAN-2, and THESAN-WC-2, in order to separately vary the resolution and the reionization history. We show such comparison in Figure 12, where we display the CDF in a selection of THESAN runs at $z = 5.5$ (top panel) and at the time when the volume-averaged hydrogen neutral fraction is $x_{\mathrm{HI}} = 0.02$ (bottom panel). In both cases we use redshift bins of width $\Delta z = 0.2$ centered on the redshift reported. For the sake of visual clarity, we do not report the range computed via bootstrap, but only the original CDFs. The top panel clearly illustrates that the effective optical depth distribution is primarily controlled by the reionization history. In fact, THESAN-WC-2 appears much more similar to THESAN-1 than THESAN-2 by virtue of having a very similar reionization history to the former and despite having the same mass resolution and force softening as the latter. The small incongruity between the weak convergence run and our flagship simulation can be entirely explained by the small residual difference in their reionization histories (see Fig. 3), as shown in the bottom panel, where all curves except THESAN-HIGH-2 and THESAN-LOW-2 perfectly overlap. It is interesting to note that – despite synchronizing the reionization histories – there are large differences in the distribution of τ_{eff} when galaxies of different masses are forced to single-handedly power the reionization process. (We have checked that this feature does not change when choosing different hydrogen
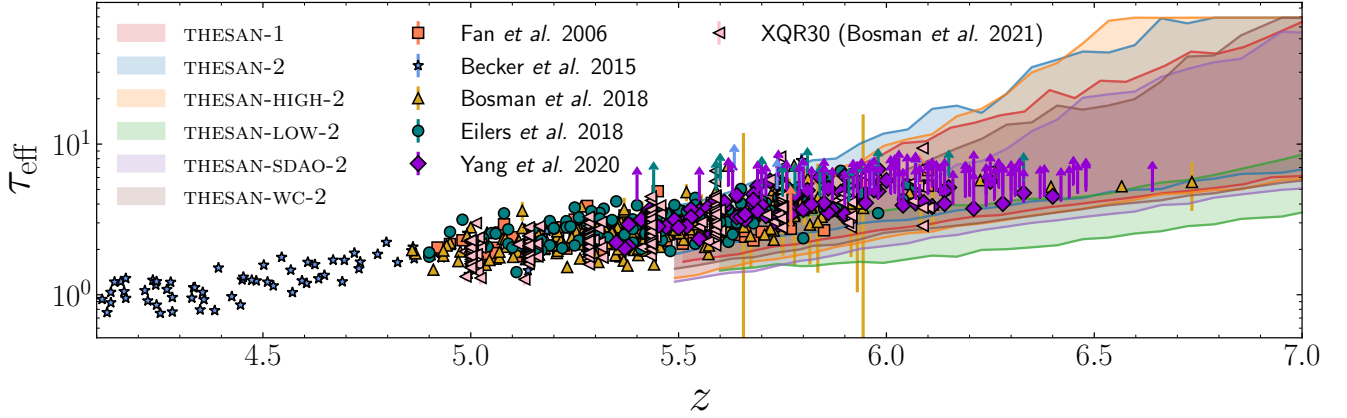
Figure 10. Evolution of the effective Ly α optical depth ($\tau_{\rm eff}$) in different redshift bins, averaged over spectral chunks of length 50 h^{-1} Mpc. The shaded regions show the central 95% of the data computed from synthetic spectra in a selection of THESAN runs, while symbols report individual observations of Fan et al. (2006), Becker et al. (2015), Bosman et al. (2018, 2021), Eilers et al. (2018), and Yang et al. (2020).

neutral fractions.) Therefore, it appears possible – at least theoretically – to combine observations of $\tau_{\rm eff}$ and $x_{\rm HI}$ to infer precious information on the sources primarily responsible for the reionization of the Universe. Additionally, the bottom panel of Figure 12 highlights again the excellent numerical convergence of the THESAN simulations with respect to the properties of the high-z IGM. Once the different reionization histories in THESAN-1, THESAN-2, and THESAN-WC-2 (due to the slightly different star-formation histories and stellar escape fractions, see Paper I) are accounted for, the properties of the IGM appear indistinguishable.

In the highest-redshift bin (i.e. $6.2 \leq z \leq 6.5$) of Figure 11 the CDF of optical depth in THESAN-1 appears significantly different from the data from Yang et al. (2020), the only one with more than a single measurement at these redshifts. This seems to imply that THESAN predicts too much inhomogeneity in the IGM, therefore requiring an earlier reionization and/or *fewer* resolved substructures. It appears very challenging to reconcile this requirement with what can be inferred at lower redshifts, unless the reionization history of the Universe is quite extended and shows little evolution between $z \sim 6.5$ and $z \sim 5.5$. Since data in this redshift range originates from a single dataset we can meaningfully compare our simulations with, we have checked that forward modelling the errors in the synthetic spectra[3] does not significantly alleviate the disagreement between THESAN-1 and the data from Yang et al. (2020).

4.2 Characterisation of transmission regions

The analysis performed in this Section focuses on the transmitted flux in the Ly α forest. Until now we have only considered its amplitude. In the following, we move to a deeper analysis, characterising the

---

[3] We include errors on the mean transmitted flux in synthetic spectra by assigning to each spectral chunk of the simulated spectra an error $\sigma_{\langle F \rangle}$ randomly extracted from the errors in the Yang et al. (2020) dataset in the same redshift bin. $\tau_{\rm eff}$ is then computed in the very same way as in observations. We additionally check the effect of shifting the value of $\langle F \rangle$ by an amount sampled from a Gaussian distribution of zero mean and width equal to $\sigma_{\langle F \rangle}$, in order to mimic the effect of systematic changes in the mean flux due to such errors. We find that this additional step does not produce any appreciable difference in the optical depth distribution at this redshift.

shape and distribution of the transmission regions. More specifically, we employ the dark gaps (dg) and peaks definition presented in Section 2.2. In Fig. 13 we show the distribution of dg lengths ($L_{\rm dg}$, left panel), peak heights ($h_{\rm p}$, central panel), and widths ($w_{\rm p}$, right panel) in the THESAN-1 simulation as a function of redshift (indicated by the line color, top row) and for the different THESAN models at the time when the volume-averaged neutral fraction is $x_{\rm HI} = 0.02$ (approximately $z \sim 5.5$ for most models, except for THESAN-LOW-2, see Fig. 3). Specifically, we show the quantity $X \, {\rm d}P(X)/{\rm d}X$ for each feature $X = \{L_{\rm dg}, h_{\rm p}, w_{\rm p}\}$, where the probability density is computed as (Gnedin et al. 2017):

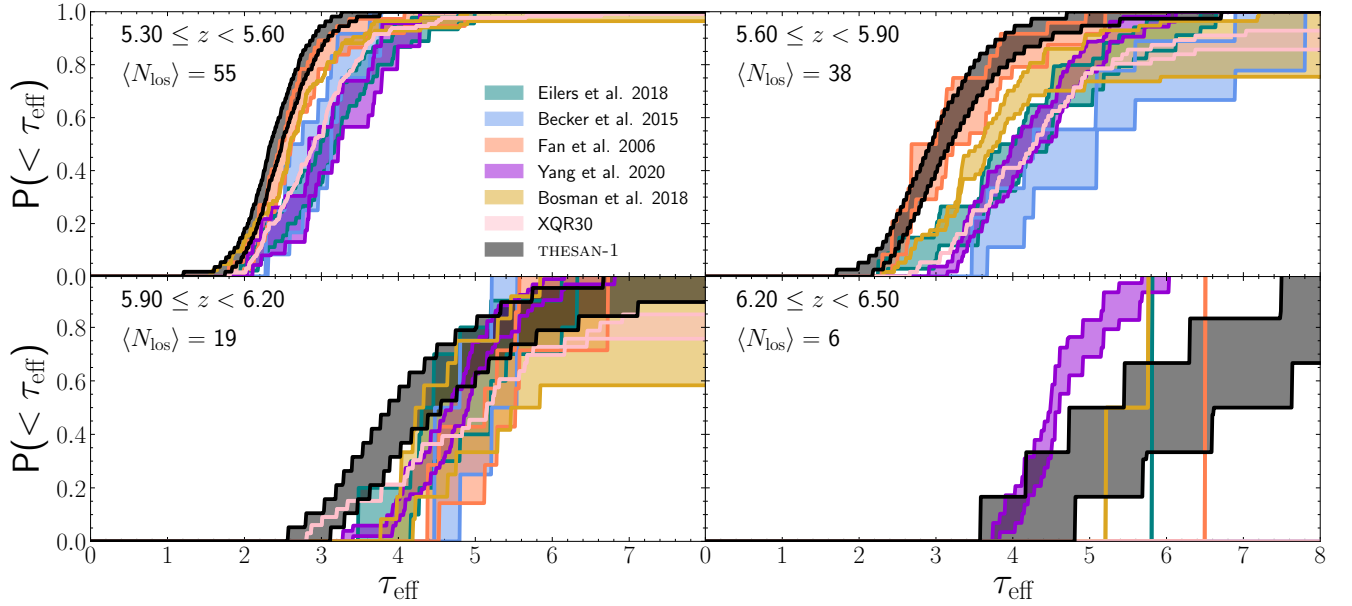$$\frac{{\rm d}P(X)}{{\rm d}X} \equiv \frac{\Delta N(X, \Delta X)}{N_{\rm tot} \Delta X} \,, \tag{3}$$

with ΔN being the number of features in a bin centred on a given value X and of width ΔX, and $N_{\rm tot}$ is the total number of features detected. It should be noted that the length of dark gaps is limited to the length of synthetic spectra, which in turn is constrained by the box length. This only becomes relevant at $z \gtrsim 6.5$, where the number of such long dark gaps is not negligible. For this analysis, we convolve the spectra with a boxcar filter that effectively degrades their spectral resolution to $R = 2000$, in order to mimic observations of these features (Gnedin et al. 2017).

A first apparent feature visible from the top panels of Fig. 13 is the redshift dependence of the distributions. The length of dark gaps (left panel) shows a flat distribution at $z \sim 7$, while at lower redshifts the number of long gaps is progressively suppressed as more highly-ionized regions emerge around the sources of radiation. By $z \sim 6$, the probability of finding dark gaps larger than 10 Mpc drops steeply. This highlights the possibility of using the tail of this distribution to constrain the timing of reionization, similarly to what has been done recently by Keating et al. (2020) with an individual sightline containing an exceptionally-long Gunn-Peterson trough. However, our results show that the occurrence of short gaps can also provide valuable information on the reionization history of the Universe, opening up the possibility of employing the full distribution to constrain the latter.

Initial steps in this direction were taken in Gnedin et al. (2017), although in that work a single reionization history was investigated. Differences in the distribution of dark gap lengths can also arise also from the different source models for reionization. Therefore, in the

Figure 11. Cumulative distribution function of effective Ly α optical depths in different redshift bins, averaged over spectral chunks of length 50 h^{-1} Mpc, and compared to the observations of Fan et al. (2006), Becker et al. (2015), Bosman et al. (2018, 2021), Eilers et al. (2018), and Yang et al. (2020). The ranges are computed by bootstrapping (with 500 samples) the original distribution and showing the 15th and 85th percentile of the samples, except for the Bosman et al. (2018) sample, where we follow their optimistic-pessimistic approach (see the text for an explanation).
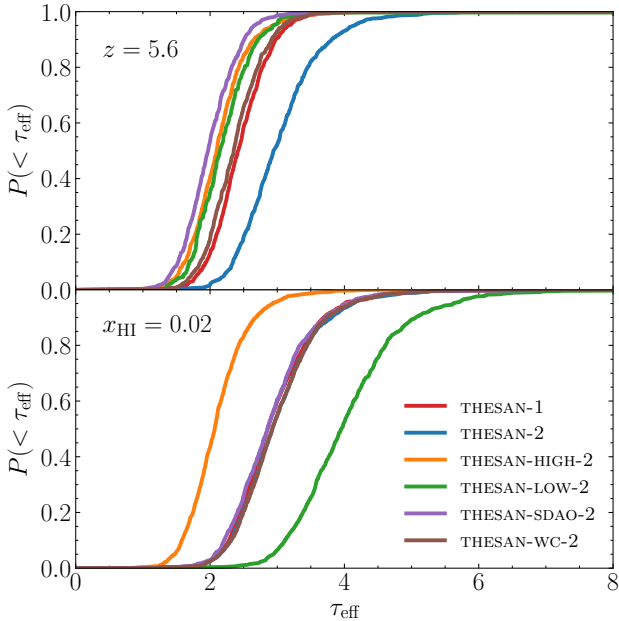
Figure 12. Cumulative distribution function of effective Ly α optical depths at $z = 5.6$ (top) and at the same volume-averaged hydrogen neutral fraction $x_{HI} = 0.02$ (bottom) for a selection of THESAN runs.

bottom left panel of the figure we show the distribution in different THESAN runs at the same volume-averaged hydrogen neutral fraction $x_{HI} = 0.02$. This removes differences arising from unequal reionization histories (although some residual differences may be found since we are now comparing distributions at slightly different red-

shifts, and therefore the density field in these runs is not exactly the same, despite originating from the same initial conditions). We have checked that the residual differences in the temperature evolution after the reionization history is controlled for do not play a major role. For instance, the evolution of $T(x_{HI})$ in THESAN-1 and THESAN-LOW-2 is practically indistinguishable, while they are well separated in the bottom panel of Fig. 12. It is however desirable that these results are checked in greater detail, especially given their potential for constraining the sources of reionization. All runs that share the same source model (THESAN-1, THESAN-2, and THESAN-WC-2) show indistinguishable distributions. Interestingly, the distribution computed from THESAN-SDAO-2 overlaps with the former ones, showing that this statistical measure is not able to distinguish ΛCDM from the sDAO dark matter model. This is likely a consequence of the fact that the sDAO model differs from the ΛCDM one only in the abundance of small dark matter haloes, therefore impacting mostly the initial phases of reionization, which are dominated by such objects (see Paper I), while becoming indistinguishable from the other runs once more massive objects start to dominate the ionizing photon budget. It should be noted, however, that the definition of dark gaps itself depends on a parameter (the flux threshold) that can change the sensitivity of $P(L_{dg})$ to different physical quantities. Finally, THESAN-LOW-2 and THESAN-HIGH-2 show a somewhat counter-intuitive behaviour. The former shows more long gaps while the latter shows fewer, despite the former harbouring more ionizing sources than the latter. This behaviour stems from the fact that the small galaxies sourcing reionization in THESAN-LOW-2 produce lower transmission with respect to THESAN-HIGH-2 and, therefore, more of these peaks remain below the flux threshold used to define a dark gap in the former, but not in the latter, effectively joining together what in THESAN-HIGH-2 are multiple adjacent gaps. This is also the reason why the number of short gaps is larger (smaller) in THESAN-HIGH-2 (THESAN-LOW-2).

It is interesting to note that the shape of $P(L_{dg})$ depends on red-
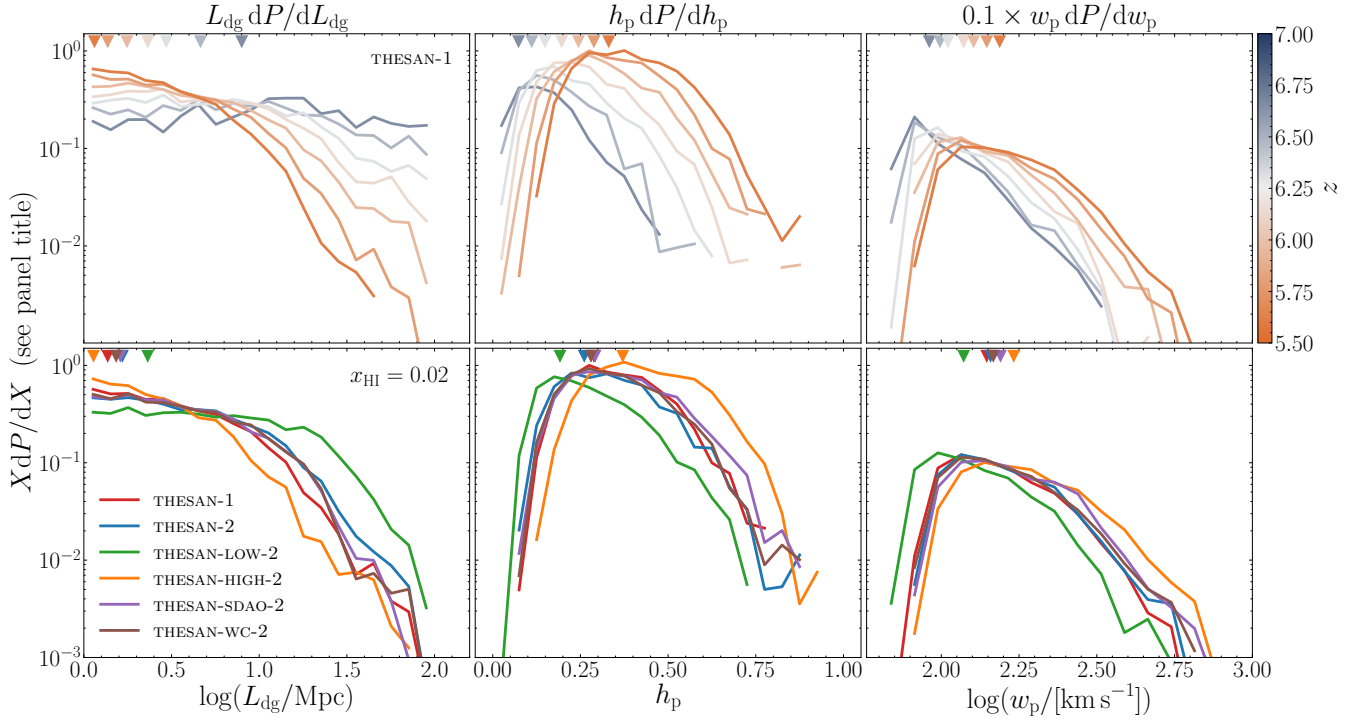
Figure 13. *Top*: Distribution of dark gaps lengths (L_{DG}, left panel), peak heights (h_p, central panel), and peak widths (w_p, right panel) in the Lyα forest as a function of redshift in the THESAN-1 run. *Bottom*: Same as above, but at fixed volume-averaged neutral fraction ($x_{HI} = 0.02$) in a selection of THESAN runs. Triangles on the top part of the panels show the location of the mean of each distribution.

shift and the reionization model (i.e. driven by low- or high-mass galaxies) in different ways. In particular, the main effect of the former is to split long dg into multiple, shorter ones, decreasing $P(L_{dg} \gtrsim 10\,\mathrm{Mpc})$ while boosting $P(L_{dg} \lesssim 5\,\mathrm{Mpc})$. Interestingly, the position of this break remains approximately fixed. Varying the source models, instead, moves the position of this break towards larger (smaller) lengths when only photons from low (high) mass galaxies reach the IGM. This is very important as it enables us to disentangle the two effects from observational data.

The behaviour described is consistent with the results presented in the central and right panels of Fig. 13. In particular, the distributions of both h_p and w_p move to larger values as reionization progresses, the IGM becomes more ionized and the ionized patches grow larger (see Garaldi et al. 2019b, for a detailed discussion of how the peak shape depends on IGM properties). Once the reionization history is controlled for, THESAN-LOW-2 exhibits peak heights that are, on average, smaller than in our fiducial physical model. Conversely, the transmission peaks in THESAN-HIGH-2 are on average higher. A similar trend can be seen in the peak widths, with the former (latter) model producing narrower (broader) peaks than in THESAN-1.

5 THE IGM – GALAXY CONNECTION

In the previous sections we have demonstrated that THESAN predicts very well the observed properties of the high-z IGM. In Paper I we have shown that our simulations produce a realistic galaxy population at $z \gtrsim 5.5$ (and we are confident the same is true at lower redshifts

since we use the successful IllustrisTNG galaxy formation model). Hence, we investigate here the connection between the high-redshift galaxies and the IGM.

We do so by using the dependence of the average transmissivity T (i.e. the ratio between the transmitted flux and the inferred continuum flux at the same location) in the Lyα forest at a distance r from nearby galaxies as a probe of the connection between galaxies and the IGM at $z \geq 5.5$. This test that was originally proposed and observationally performed by Kakiichi et al. (2018). Subsequently, additional observations, as well as a refinement of the approach itself, have been provided by Meyer et al. (2019) and Meyer et al. (2020). In particular, Meyer et al. (2019) have shown that at $4.5 \lesssim z \lesssim 6.3$ the transmitted flux displays a broad peak at intermediate distances to galaxies $10 \lesssim r/[\mathrm{cMpc}/h] \lesssim 30$, which has been interpreted as a proximity effect of the latter. In fact, under the assumption that galaxies provide the main contribution to the reionization process, their surroundings are expected to experience a larger-than-average density of ionizing photons, which entails a larger IGM transmission.[4]

On the theoretical side, however, this approach has been barely investigated, potentially because it simultaneously requires large simulated volumes and radiative transfer techniques to faithfully capture

---

[4] Notice that, even in the case QSOs were the dominant source of ionizing photons (which is not the case in THESAN, see Paper I), we expect the aforementioned flux excess to exist. In fact, QSOs are co-spatial with a subset of galaxies, therefore producing a flux enhancement around the latter. However, since QSOs are significantly more rare, we also expect the signal excess to be washed out when computing the correlation using all galaxies.
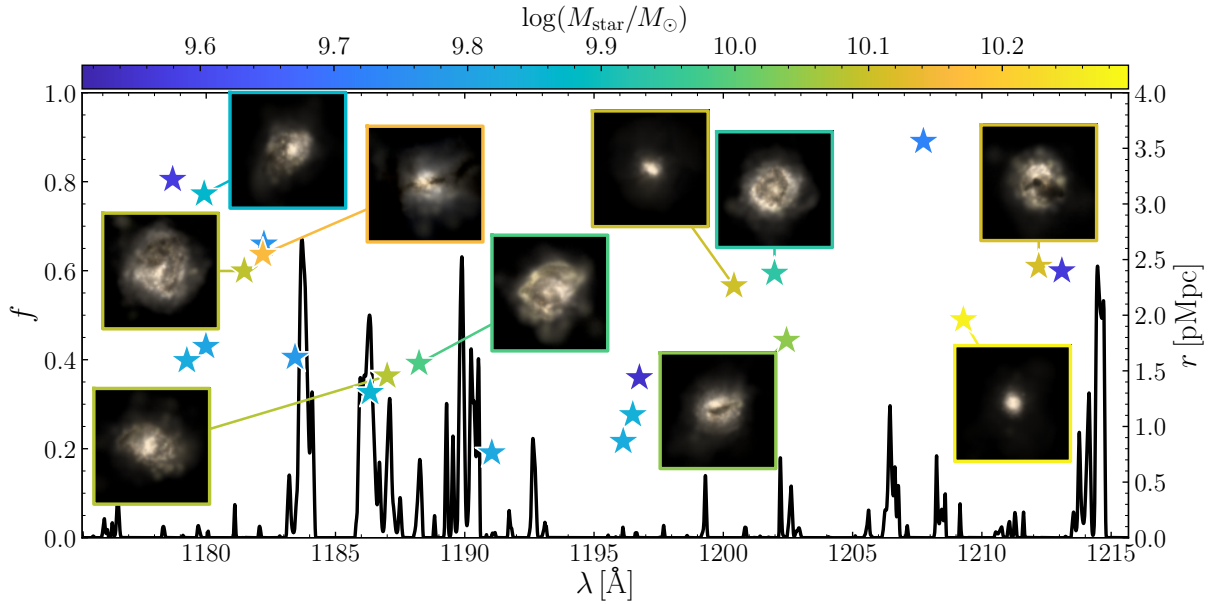
Figure 14. Normalised Ly α transmitted flux along a sightline through the THESAN-1 box (left vertical scale). Star symbols show the surrounding galaxies with stellar mass $M_{\rm star} \geq 10^{9.5}\,M_{\odot}$, colour-coded according to the latter. Their vertical position reflects their distance from the sightline (right vertical scale). Insets show synthetic JWST mock images of the ten galaxies in the panel with the largest $M_{\rm star}$, combining the F277W, F356W and F444W NIRCam wide filters and covering an area of $(10\,{\rm pkpc})^2$.

the IGM properties, as well as a realistic galaxy formation model, which in turn often requires high resolution. For these reasons, THESAN is ideally suited to study this quantity. Garaldi et al. (2019b) have investigated the ability of the state-of-the-art Cosmic Reionization On Computer simulations (Gnedin 2014) to reproduce the observed modulation of $T(r)$, and concluded that the excess transmission at intermediate distances observed is not found in the CROC suite. Although the reason for this mismatch appears to be a too homogeneous radiation field in the simulated IGM, its root cause could not be firmly identified because the multiple explanations proposed cannot be extensively tested using expensive radiation-hydrodynamic simulations. Among the proposed sources of the mismatch in the simulations, we mention a too low mass for the smallest galaxies contributing significantly to the ionizing photons budget, an earlier reionization history, and a large stellar escape fraction. Additionally, the stellar-to-halo-mass relation in CROC is very flat at halo masses $M_{\rm halo} \gtrsim 10^{10}\,M_{\odot}$, in disagreement with observations and theoretical models (see figure 9 in Paper I and Zhu et al. 2020), suggesting some issues with the simulated galaxy population.

We begin providing a synthetic version of the observations of Kakiichi et al. (2018), Meyer et al. (2019) and Meyer et al. (2020) in Fig. 14, where we show the transmitted (normalized) flux at $z = 5.5$ in a random sightline through the THESAN-1 box (solid black line and left vertical scale). Galaxies around the latter are shown whenever the galaxy stellar mass exceeds $M_{\rm star} \geq 10^{9.5}\,M_{\odot}$ using star symbol coloured to reflect $M_{\rm star}$ and vertically positioned to show their distance from the sightline (right vertical scale). Finally, we show mock JWST images for the ten galaxies in the plot with the largest $M_{\rm star}$. Notice that these are *not necessarily* the largest galaxies in the simulation. The JWST mock images are generated using the code SKIRT (last described in Camps & Baes 2020) and following the procedure described in Vogelsberger et al. (2020b). In particular, we show a composite image obtained combining the F277W, F356W and F444W NIRCam wide filters covering an area of $(10\,{\rm pkpc})^2$.

Moving to a quantitative analysis, we show in Fig. 15 the excess IGM transmissivity with respect to its average across all r, i.e. $\langle T(r) \rangle / \bar{T} - 1$, at $z = 5.5$ in THESAN-1. The comparison with observations is rendered difficult by the diverse observational techniques required to collect this type of data. For instance, Meyer et al. (2019, yellow squares in the figure) employed the C IV absorption along the same QSO lines of sight to estimate the position of galaxies nearby, while Meyer et al. (2020, cyan triangles in the figure) employed LAE and Lyman-break galaxies. Therefore, in an effort to enable a general comparison, we show in the figure the results obtained from THESAN-1 applying a number of different criteria in the galaxy identification. In particular, in the top panel of the Figure we report the curves corresponding to three minimum halo masses (black to grey curves), while in the middle panel we explore the effect of a minimum galaxy stellar mass (purple curves). In the bottom panel, we present the outcome of forward modelling the C IV absorption features (employing the same approach used for Lyα and described in Section 2) and using them to estimate the galaxy position along the line of sight. The orange-to-red curves report the obtained $T(r)$ for different values of the minimum C IV column density employed in the galaxy identification, namely $\log(N_{\rm CIV}/{\rm cm}^{-2}) \geq 12.5, 13, 13.5, 14$. The first value corresponds to the nominal lower threshold imposed in Meyer et al. (2019), while the last corresponds to the $N_{\rm CIV}$ for which 90% completeness is reached.

From the top panel of Fig. 15, the effect of selecting galaxies hosted in haloes of different mass can be seen. In particular, we can partially match the observed flux modulation from Meyer et al. (2019) only when selecting haloes with mass $M_{\rm halo} \gtrsim 10^{12}\,M_{\odot}$, although the signal becomes very noisy because of the small number of haloes with such mass at $z = 5.5$ in our simulation box. A better agreement is found when we instead select only galaxies with stellar mass $M_{\rm star} \geq 10^{10}\,M_{\odot}$, as seen in the middle panel, although the recovered $\langle T(r) \rangle$ from simulations is still lower than the data from Meyer et al. (2019). Consistently with the top panel, lowering the
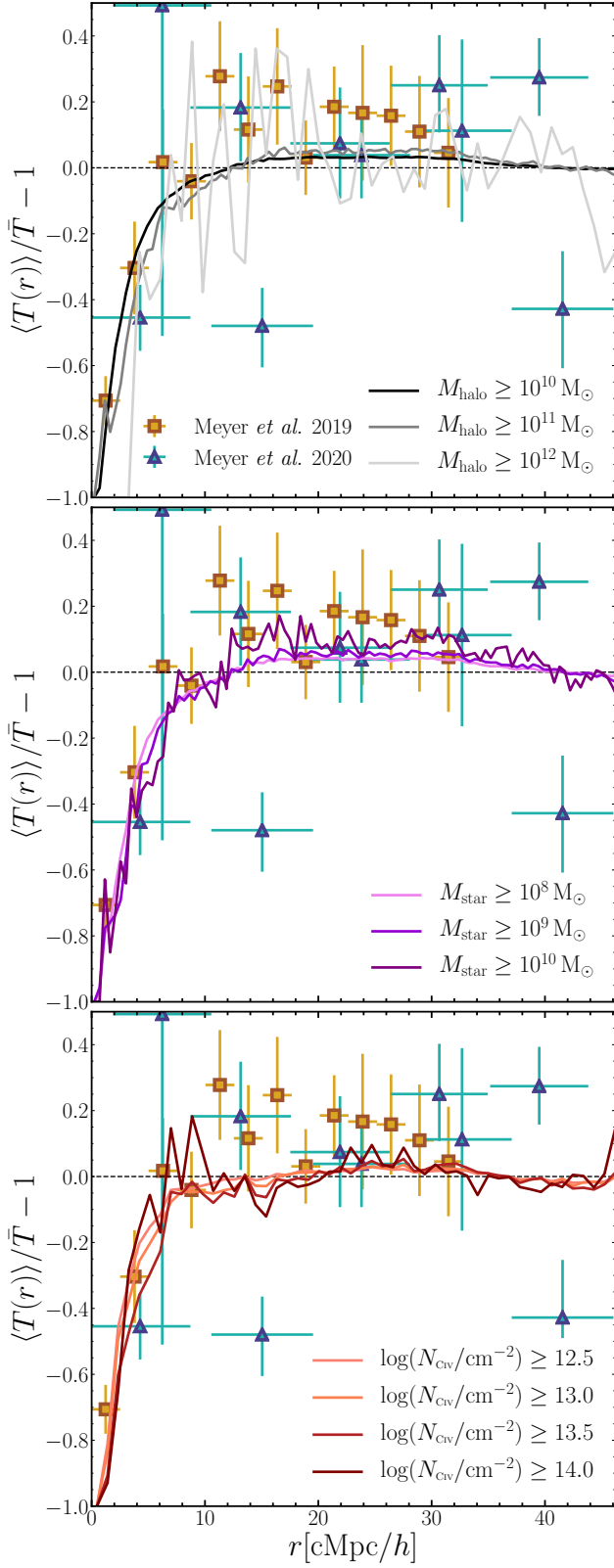
Figure 15. Average transmission in the Lyα forest in the THESAN-1 run as a function of distance from nearby galaxies at $z = 5.5$. Lines show the effect of considering only a sub-population of halos or galaxies while computing this quantity. Symbols show the measurements from Meyer et al. (2019, yellow squares) and Meyer et al. (2020, cyan triangles).
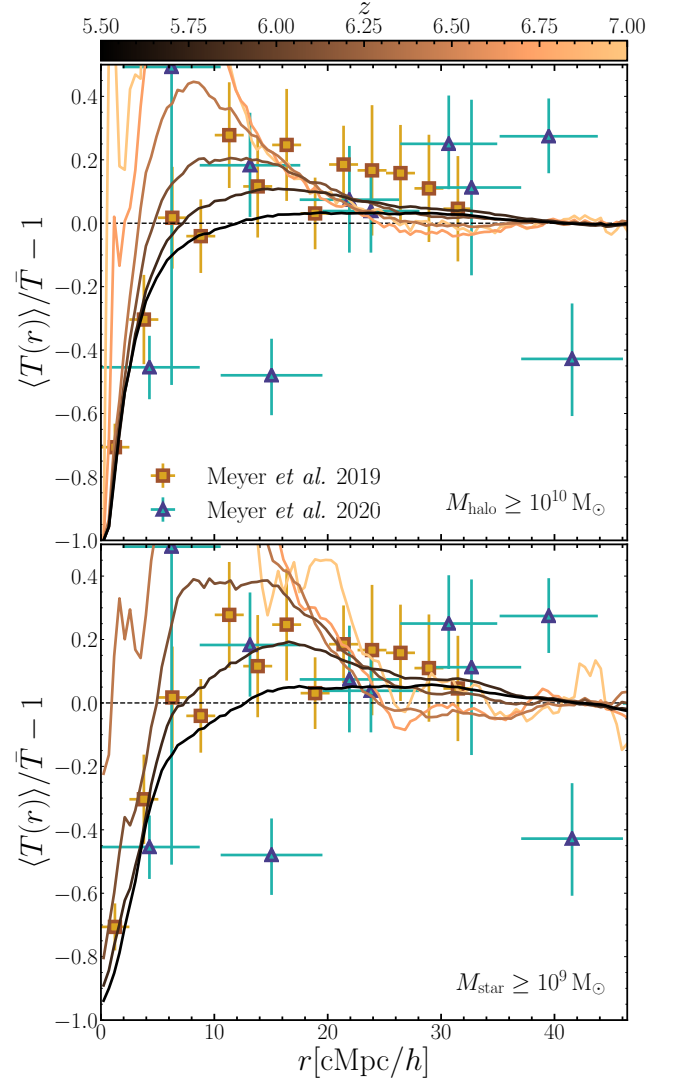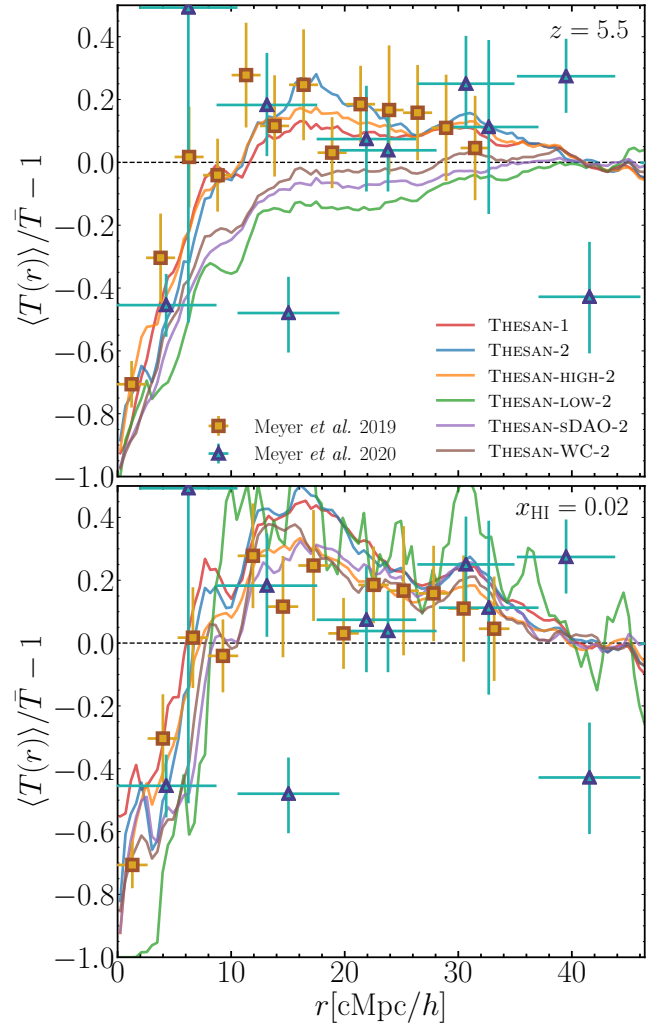

Figure 16. As Fig. 15, but now showing the redshift evolution for a single selection criterion of $M_{\rm halo} \geq 10^{10} \, M_\odot$ for the top panel and $M_{\rm star} \geq 10^9 \, M_\odot$ for the bottom one.

threshold in the stellar mass of selected galaxies (which effectively corresponds to lowering their host halo mass) reduces the excess flux at intermediate scales. This result highlights the importance of a solid characterisation of the galaxies employed to compute this correlation in order to faithfully interpret the results. While such characterisation can be achieved when surveys are employed, identifying nearby galaxies from absorption features in the same spectrum renders the task significantly more challenging. Finally, when employing the synthetic C IV absorption to locate galaxies along the line-of-sight (bottom panel of Fig. 15), we fail to reproduce the observed transmissivity modulation for all the column density thresholds employed (note that in THESAN-1 at $z = 5.5$ there are no absorption features with $\log(N_{\rm CIV}/{\rm cm}^{-2}) \geq 14.5$). Before investigating the reason for this failure in the following, we note here that the C III ionization frequency is just slightly higher than the ionization frequency of He I. Hence, the suppression of the radiation field at such frequencies due to helium (double) reionization is highly dependent on both the details of the latter in the simulations. In particular, the radiation field

intensity at the C IV ionization frequency in the UVBG employed (i.e. Faucher-Giguère et al. 2009, see Sec. 2) varies by 40 orders of magnitude in the redshift range $6 \leq z \leq 10$ (as a consequence of the evolving absorption due to helium secondary ionization. Therefore, even a small difference in the reionization history assumed in the UVBG can have catastrophic consequences for the simulated abundance of C IV. We plan to quantify this effect better, and improve upon our current method, in a future work, where we will also investigate the feasibility of using the observed C IV line evolution to constrain high-z helium reionization (and, consequently, the coeval UVBG and sources of energetic photons).

In order to investigate the reasons for this failure, we start by showing in Fig. 16 the redshift evolution (as indicated by the line colour) of $T(r)$. For the sake of clarity, we only show two selection criteria, namely $M_{\rm halo} \geq 10^{10}\,{\rm M}_\odot$ and $M_{\rm star} \geq 10^9\,{\rm M}_\odot$. This choice is made to ensure that a sufficient number of galaxies is selected at each redshift. The curves in both panels show the same evolution, where two main effects can be appreciated. On the one hand, the signal appears stronger at higher redshift, reflecting how the average transmissivity $\bar{T}$ decreases at earlier epochs (see e.g. Fig. 9 for the actual mean flux evolution) much more rapidly than the transmission in the first highly-ionized regions producing the transmission spikes. On the other hand, the location of the excess in the transmission moves closer to the galaxies with increasing redshift. This results from the combination of two separate factors: (i) the lower overdensities in which galaxies reside at earlier times limits the suppression of the Lyα flux (because of enhanced hydrogen recombination) at smaller and smaller distances from the galaxy itself; and (ii) the smaller sizes of ionized bubbles (see e.g. Fig. 13 or fig. 19 in Paper I), which reduce the distance at which the flux is enhanced. Accepting at face value the minimum C IV host mass quoted in Meyer et al. (2019), i.e. $M_{\rm halo} \geq 10^{10}\,{\rm M}_\odot$, corresponding to the top panel of Fig. 16, but see their figure 2 and related discussion about the effective halo mass traced), we find that the match between their data and the IGM properties in THESAN is improved if we employ simulation outputs at $z \sim 6.2$ in place of their mid redshift of ~ 5.5, potentially indicating the necessity for an even later reionization in our model. However, a number of other explanations are possible (see Section 1 and Garaldi et al. 2019b, for a discussion). We next investigate the effects of the nature of dark matter and of the escape of photons from galaxies using the THESAN medium resolution runs.

In Fig. 17 we show the results obtained in a selection of THESAN runs. We choose to show the curves computed selecting only galaxies with $M_{\rm star} \geq 10^{10}\,{\rm M}_\odot$, as this is the one that best matches the observed data at $z = 5.5$ in THESAN-1. We show the results from different runs at the same redshift (top panel) and neutral hydrogen fraction (bottom panel). It is immediately clear that, when a single redshift is selected, the different runs exhibit two very different types of behaviours. On one hand THESAN-1, THESAN-2 and THESAN-HIGH-2 agree fairly well with the observed modulation, while THESAN-LOW-2, THESAN-SDAO-2 and THESAN-WC-2 do not exhibit any enhancement of transmission at intermediate scales, scoring very poorly in their comparison with data. Once we match the ionization state of the IGM, however, all curves appear very similar. While this is easily understandable for most of the models, it may be surprising that the same occurs for THESAN-LOW-2, where one could expect the signal to be reduced since only large galaxies are selected. Nevertheless, the clustering of small objects around larger ones produces a large-scale reionization topology which is similar to the one in the fiducial model in the vicinity of massive objects. This can be visually appreciated in Fig. 1, where the majority of differences between THESAN-2 and THESAN-LOW-2 are found around small, isolated galaxies. Therefore,

Figure 17. As Fig. 15, but now showing the impact of high-z physics at fixed redshift ($z = 5.5$, top panel) and volume-averaged hydrogen neutral fraction ($x_{\rm HI} = 0.02$, bottom).

$T(r)$ appears to be robust against the physical model differences we have explored in THESAN, making it a promising tool for constraining the timing of reionization. Comparing the differences between THESAN-1 and THESAN-WC-2 in the top panel of Fig. 17 to their small difference in reionization histories (see Fig. 3) it can be inferred that $T(r)$ is extremely sensitive to the reionization history of the Universe, rendering it a powerful probe of the latter, but also highlighting the necessity of more detailed studies. Unfortunately, the nature of this probe – at the interface between reionization and galaxy formation and at scales where radiation feedback can play a relevant role – renders its study significantly expensive from a computational perspective.

6 SUMMARY AND CONCLUSIONS

We have introduced the THESAN project, a suite of radiation-magneto-hydrodynamic simulations designed to simultaneously and self-consistently model the Epoch of Reionization and the formation of the first galaxies. The simulations have box sizes of $L_{\rm box} = 95.5\,{\rm Mpc}$,

resolve atomic cooling haloes, and employ a well-tested galaxy formation model (the same developed for the IllustrisTNG suite), augmented with self-consistent radiation transport, a non-equilibrium thermo-chemistry solver, and dust creation/destruction processes. The different runs explore changes in the escape of ionizing photons from galaxies and in dark matter properties. This paper is part of a series of three introductory manuscripts (alongside with Paper I and Paper II), and focuses on a thorough analysis of the properties of the high-redshift intergalactic medium, its connection with the galaxy population, and Lyα transmission through the IGM. We summarise our main results in the following points:

(i) The THESAN simulations have realistic reionization histories, matching most of the available constraints on the evolution of the global neutral hydrogen fraction. They all follow a 'late reionization' model, completing at $z \lesssim 5.5$, with the exception of THESAN-LOW-2, where only small galaxies emit ionizing photons and the IGM is completely ionized by $z \sim 6.5$. The gas temperature at mean density and the optical depth of CMB photons have realistic values in all runs (see Fig. 3).

(ii) The low-density warm gas in THESAN shows a realistic distribution of densities and temperatures thanks to the self-consistent treatment of radiation. Its distribution (Fig. 4) reflects the locally-different reionization histories (see also Fig. 1) of gas regions exposed to different radiation fields. Consequently, the heat injection from photo-ionization occurs at different times in the universe, and therefore gas patches at a given redshift experienced unequal cooling, producing a wide range of temperatures.

(iii) The number of photons produced by galaxies (as measured by the ionizing emissivity, Fig. 6) appears to be only slightly larger, if extrapolated to $z \leq 5$, than in available observations. The predicted evolution does not show any sign of a drop, as assumed in some recent semi-numerical studies (e.g. Kulkarni et al. 2019a; Keating et al. 2020, although we postpone to a future work the investigation of the effect of an evolving ionizing photon escape fraction from haloes). Some of these photons make their way to the ionized IGM patches, building up the photo-ionization rate (Fig. 5), which appears in good agreement with observed values. The main exception is THESAN-LOW-2, which significantly over-estimates the measurements at $z \geq 5.5$.

(iv) The properties of the Lyα forest in most THESAN runs match the observed mean flux (Fig. 9) and optical depth evolution (Fig. 10) very well. When we force only small galaxies to produce ionizing photons, the mean flux is significantly higher than observed (and, consequently, the effective optical depths significantly lower) as a consequence of the early reionization history of this run.

(v) The IGM in THESAN-1 appears slightly too homogeneous, as shown by the distribution of optical depths (Fig. 11), which is only marginally consistent with available observations. This may be evidence in favour of a reionization history completing later than in our model (i.e. at $z \lesssim 5.2$), as can be seen by looking at the same observable in THESAN-2 (Fig. 12), which has such delayed reionization with respect to the THESAN-1 run.

(vi) The mean free path of ionizing photons (Fig. 7) shows rapid evolution between $z = 5$ and $z = 6$ in all runs except THESAN-LOW-2. The values computed from our main run (THESAN-1) align very well with recent measurements.

(vii) We characterise the shape of high-z transmission regions in the Lyα forest as well as the low-flux gaps between them (Fig. 13). We find they evolve significantly with redshift, opening up the possibility of constraining the history of reionization using this statistic. Additionally, we find that these measurements are sensitive to the mass of the sources powering the EoR (i.e. they are able to distinguish a

scenario where small galaxies produce the bulk of ionizing photons from one where large ones power the reionization of hydrogen).

(viii) We study the modulation of the transmitted Lyα flux as a function of the distance from galaxies (Fig. 15). This is a particularly stringent test of the simulations, since it depends simultaneously on the properties of reionization and of the galaxy population. In the past even state-of-the-art simulations could not reproduce the observed flux enhancement at intermediate scales (Garaldi et al. 2019b). We find that we can reproduce this general observed trend, but not its amplitude, which is smaller in THESAN than in observed QSO spectra (Meyer et al. 2019). We investigate possible reasons for this difference and find that, consistently with other results presented, a later reionization would bring the simulation into better agreement with observed values (Fig. 16). (Note however that the latter come from only a handful of sightlines.) Finally, we find that this measure is robust against the detailed modelling of the sources of reionization (Fig. 17); i.e. a similar flux modulation is produced for all the models investigated, once their reionization history is accounted for.

In this work we have presented the properties of the high-z intergalactic medium and Lyα transmission using the new THESAN simulation suite. We find an excellent agreement between the simulated properties and available observations, despite not explicitly tuning parameters to achieve such a result. The THESAN simulations represent the first radiation-hydrodynamic simulations of the 'late reionization' model, which was recently suggested to explain the long Gunn-Peterson trough observed. We show that this model is also needed to account for the rapid evolution of the ionizing photon mean free path and effective optical depth distributions. THESAN improves on the existing literature on this subject by: including a significantly-expanded range of physical processes, reaching higher resolution, and self-consistently following the propagation of radiation on-the-fly. Additionally, we provide a thorough analysis of the connection between the properties of reionization and the galaxy population at $z \gtrsim 5.5$, which is one of the biggest challenges that theoretical modelling will face as the properties of the first galaxies become the new frontier in studies of the EoR and galaxy formation, thanks to the flood of information that will be provided by instruments like JWST, ALMA, HERA, SKA, CCAT-p and SPHEREx. We speculate that an even-later reionization history would likely cure most of the small discrepancies with observations.

The simulations introduced and the forward modelling framework presented here open up many exciting prospects for the investigation of the high-z Universe. A possible avenue forward, which we plan to pursue in a future work, is to employ simulations to decode observations of high-redshift metal lines in the spectra of background QSOs. These features have the potential to unveil the production and distribution of the first metals in the Universe by the first galaxies, therefore complementing the analysis presented here, which has focused on the diffuse hydrogen. Additionally, the metal content of the IGM and CGM is another interesting research path, since recent observational campaigns have shown that virtually all metals reside in the cold gas phase at $z \gtrsim 2.5$ (Péroux & Howk 2020).

In conclusion, the results described in this and in the accompanying papers provide a solid base for advancing our knowledge of cosmic reionization and high-z structure formation using the THESAN simulations, which we will make public in the near future.

DATA AVAILABILITY

All simulation data, including snapshots, group and subhalo catalogues, merger trees, and high time cadence Cartesian outputs will be made publicly available in the near future. Data will be distributed via `www.thesan-project.com`. Before the public data release, data underlying this article will be shared on reasonable request to the corresponding author(s).

REFERENCES

Atek H., Richard J., Kneib J.-P., Schaerer D., 2018, MNRAS, 479, 5184
Barnes J., Hut P., 1986, Nature, 324, 446
Barnett R., Warren S. J., Becker G. D., Mortlock D. J., Hewett P. C., McMahon R. G., Simpson C., Venemans B. P., 2017, A&A, 601, A16
Bauer A., Springel V., Vogelsberger M., Genel S., Torrey P., Sijacki D., Nelson D., Hernquist L., 2015, MNRAS, 453, 3593
Becker G. D., Bolton J. S., 2013, MNRAS, 436, 1023
Becker R. H., et al., 2001, AJ, 122, 2850
Becker G. D., Bolton J. S., Madau P., Pettini M., Ryan-Weber E. V., Venemans B. P., 2015, MNRAS, 447, 3402
Becker G. D., D'Aloisio A., Christenson H. M., Zhu Y., Worseck G., Bolton J. S., 2021, arXiv e-prints, p. arXiv:2103.16610
Bird S., Haehnelt M., Neeleman M., Genel S., Vogelsberger M., Hernquist L., 2015, MNRAS, 447, 1834
Bohr S., Zavala J., Cyr-Racine F.-Y., Vogelsberger M., Bringmann T., Pfrommer C., 2020, MNRAS, 498, 3403
Bolton J. S., Becker G. D., Wyithe J. S. B., Haehnelt M. G., Sargent W. L. W., 2010, MNRAS, 406, 612
Bolton J. S., Becker G. D., Raskutti S., Wyithe J. S. B., Haehnelt M. G., Sargent W. L. W., 2012, MNRAS, 419, 2880
Bondi H., 1952, MNRAS, 112, 195
Bondi H., Hoyle F., 1944, MNRAS, 104, 273
Bosman S. E. I., 2021, arXiv e-prints, p. arXiv:2108.12446
Bosman S. E. I., Fan X., Jiang L., Reed S., Matsuoka Y., Becker G., Haehnelt M., 2018, MNRAS, 479, 1055
Bosman S. E. I., et al., 2021, arXiv e-prints, p. arXiv:2108.03699
Bouwens R. J., Illingworth G. D., Oesch P. A., Caruana J., Holwerda B., Smit R., Wilkins S., 2015, ApJ, 811, 140
Bouwens R. J., et al., 2021, arXiv e-prints, p. arXiv:2106.13719

Cabral B., Leedom L. C., 1993, in Proceedings of the 20th annual conference on Computer graphics and interactive techniques. pp 263–270
Cain C., D'Aloisio A., Gangolli N., Becker G. D., 2021, arXiv e-prints, p. arXiv:2105.10511
Calverley A. P., Becker G. D., Haehnelt M. G., Bolton J. S., 2011, MNRAS, 412, 2543
Camps P., Baes M., 2020, Astronomy and Computing, 31, 100381
Chabrier G., 2003, ApJ, 586, L133
Chornock R., Berger E., Fox D. B., Lunnan R., Drout M. R., Fong W.-f., Laskar T., Roth K. C., 2013, ApJ, 774, 26
Choudhury T. R., Puchwein E., Haehnelt M. G., Bolton J. S., 2015, MNRAS, 452, 261
Croton D. J., et al., 2016, ApJS, 222, 22
D'Aloisio A., McQuinn M., Davies F. B., Furlanetto S. R., 2018, MNRAS, 473, 560
D'Aloisio A., McQuinn M., Maupin O., Davies F. B., Trac H., Fuller S., Upton Sanderbeck P. R., 2019, ApJ, 874, 154
Davies F. B., Furlanetto S. R., 2016, MNRAS, 460, 1328
Davies F. B., et al., 2018, ApJ, 864, 142
Davies F. B., Bosman S. E. I., Furlanetto S. R., Becker G. D., D'Aloisio A., 2021, arXiv e-prints, p. arXiv:2105.10518
Deparis N., Aubert D., Ocvirk P., Chardin J., Lewis J., 2019, A&A, 622, A142
Dijkstra M., Mesinger A., Wyithe J. S. B., 2011, MNRAS, 414, 2139
Eilers A.-C., Davies F. B., Hennawi J. F., 2018, ApJ, 864, 53
Eilers A.-C., Hennawi J. F., Davies F. B., Oñorbe J., 2019, ApJ, 881, 23
Eldridge J. J., Stanway E. R., Xiao L., McClelland L. A. S., Taylor G., Ng M., Greis S. M. L., Bray J. C., 2017, Publ. Astron. Soc. Australia, 34, e058
Fan X., et al., 2006, AJ, 132, 117
Faucher-Giguère C.-A., 2020, MNRAS, 493, 1614
Faucher-Giguère C.-A., Lidz A., Zaldarriaga M., Hernquist L., 2009, ApJ, 703, 1416
Ferland G. J., et al., 2017, Rev. Mex. Astron. Astrofis., 53, 385
Furlanetto S. R., Oh S. P., 2009, The Astrophysical Journal, 701, 94
Gaikwad P., et al., 2020, MNRAS, 494, 5091
Gallerani S., Choudhury T. R., Ferrara A., 2006, MNRAS, 370, 1401
Gallerani S., Ferrara A., Fan X., Choudhury T. R., 2008, MNRAS, 386, 359
Garaldi E., Compostella M., Porciani C., 2019a, MNRAS, 483, 5301
Garaldi E., Gnedin N. Y., Madau P., 2019b, ApJ, 876, 31
Gnedin N. Y., 2000, ApJ, 542, 535
Gnedin N. Y., 2014, ApJ, 793, 29
Gnedin N. Y., 2016, ApJ, 833, 66
Gnedin N. Y., Becker G. D., Fan X., 2017, ApJ, 841, 26
Grand R. J. J., et al., 2017, MNRAS, 467, 179
Greig B., Mesinger A., Haiman Z., Simcoe R. A., 2017, MNRAS, 466, 4239
Greig B., Mesinger A., Bañados E., 2019, MNRAS, 484, 5094
Grevesse N., Asplund M., Sauval A. J., Scott P., 2010, Ap&SS, 328, 179
Haardt F., Madau P., 1996, ApJ, 461, 20
Haardt F., Madau P., 2012, ApJ, 746, 125
Haiman Z., 2016, Cosmic Reionization and the First Nonlinear Structures in the Universe. p. 1, doi:10.1007/978-3-319-21957-8_1
Harris Daniel L. I., 1948, ApJ, 108, 112
Hassan S., Davé R., McQuinn M., Somerville R. S., Keating L. C., Anglés-Alcázar D., Villaescusa-Navarro F., Spergel D. N., 2021, arXiv e-prints, p. arXiv:2109.03840
Hjerting F., 1938, ApJ, 88, 508
Hoag A., et al., 2019, ApJ, 878, 12
Hunter J. D., 2007, Computing In Science & Engineering, 9, 90
Hutter A., 2018, MNRAS, 477, 1549
Jensen H., Laursen P., Mellema G., Iliev I. T., Sommer-Larsen J., Shapiro P. R., 2013, MNRAS, 428, 1366
Jones E., Oliphant T., Peterson P., et al., 2001, SciPy: Open source scientific tools for Python, http://www.scipy.org/
Jung I., et al., 2020, ApJ, 904, 144
Kakiichi K., et al., 2018, MNRAS, 479, 43
Kannan R., Vogelsberger M., Marinacci F., McKinnon R., Pakmor R., Springel V., 2019, MNRAS, 485, 117

Kannan R., Garaldi E., Smith A., Pakmor R., Springel V., Vogelsberger M., Hernquist L., 2021, arXiv e-prints, p. arXiv:2110.00584

Katz H., et al., 2019, MNRAS, 487, 5902

Keating L. C., Puchwein E., Haehnelt M. G., 2018, Monthly Notices of the Royal Astronomical Society, 477, 5501

Keating L. C., Weinberger L. H., Kulkarni G., Haehnelt M. G., Chardin J., Aubert D., 2020, MNRAS, 491, 1736

Kulkarni G., Choudhury T. R., Puchwein E., Haehnelt M. G., 2017, MNRAS, 469, 4283

Kulkarni G., Keating L. C., Haehnelt M. G., Bosman S. E. I., Puchwein E., Chardin J., Aubert D., 2019a, MNRAS, 485, L24

Kulkarni G., Worseck G., Hennawi J. F., 2019b, MNRAS, 488, 1035

Le Fèvre O., et al., 2020, A&A, 643, A1

Levermore C. D., 1984, J. Quant. Spectrosc. Radiative Transfer, 31, 149

Livermore R. C., Finkelstein S. L., Lotz J. M., 2017, ApJ, 835, 113

Loeb A., Barkana R., 2001, ARA&A, 39, 19

Lu T.-Y., et al., 2020, ApJ, 893, 69

Lusso E., Worseck G., Hennawi J. F., Prochaska J. X., Vignali C., Stern J., O'Meara J. M., 2015, MNRAS, 449, 4204

Madau P., Haardt F., 2015, ApJ, 813, L8

Madau P., Haardt F., Rees M. J., 1999, ApJ, 514, 648

Marinacci F., et al., 2018, MNRAS, 480, 5113

Mason C. A., Treu T., Dijkstra M., Mesinger A., Trenti M., Pentericci L., de Barros S., Vanzella E., 2018, ApJ, 856, 2

Mason C. A., et al., 2019, MNRAS, 485, 3947

McGreer I. D., Mesinger A., Fan X., 2011, MNRAS, 415, 3237

McGreer I. D., Mesinger A., D'Odorico V., 2015, MNRAS, 447, 499

McKinnon R., Torrey P., Vogelsberger M., 2016, MNRAS, 457, 3775

McKinnon R., Torrey P., Vogelsberger M., Hayward C. C., Marinacci F., 2017, MNRAS, 468, 1505

Mesinger A., Aykutalp A., Vanzella E., Pentericci L., Ferrara A., Dijkstra M., 2014, MNRAS, 446, 566

Meyer R. A., Bosman S. E. I., Kakiichi K., Ellis R. S., 2019, MNRAS, 483, 19

Meyer R. A., et al., 2020, MNRAS, 494, 1560

Mirocha J., Skory S., Burns J. O., Wise J. H., 2012, ApJ, 756, 94

Mitchell P. D., Blaizot J., Cadiou C., Dubois Y., Garel T., Rosdahl J., 2021, MNRAS, 501, 5757

Mortlock D. J., et al., 2011, Nature, 474, 616

Mutch S. J., Geil P. M., Poole G. B., Angel P. W., Duffy A. R., Mesinger A., Wyithe J. S. B., 2016, MNRAS, 462, 250

Naidu R. P., Tacchella S., Mason C. A., Bose S., Oesch P. A., Conroy C., 2020, ApJ, 892, 109

Naiman J. P., et al., 2018, MNRAS, 477, 1206

Nelson D., et al., 2018a, MNRAS, 475, 624

Nelson D., et al., 2018b, MNRAS, 477, 450

Nelson D., et al., 2019a, Computational Astrophysics and Cosmology, 6, 2

Nelson D., et al., 2019b, MNRAS, 490, 3234

Ocvirk P., et al., 2016, MNRAS, 463, 1462

Ocvirk P., Aubert D., Chardin J., Deparis N., Lewis J., 2019, A&A, 626, A77

Ocvirk P., et al., 2020, MNRAS,

Ocvirk P., Lewis J. S. W., Gillet N., Chardin J., Aubert D., Deparis N., Thelie E., 2021, arXiv e-prints, p. arXiv:2105.01663

Okamoto T., Gao L., Theuns T., 2008, MNRAS, 390, 920

Ono Y., et al., 2012, ApJ, 744, 83

Ota K., et al., 2008, ApJ, 677, 12

Ota K., et al., 2010, ApJ, 722, 803

Ouchi M., et al., 2010, ApJ, 723, 869

Pagano L., Delouis J. M., Mottet S., Puget J. L., Vibert L., 2020, A&A, 635, A99

Pallottini A., Ferrara A., Gallerani S., Vallini L., Maiolino R., Salvadori S., 2017, MNRAS, 465, 2540

Paschos P., Norman M. L., 2005, ApJ, 631, 59

Pawlik A. H., Rahmati A., Schaye J., Jeon M., Dalla Vecchia C., 2017, MNRAS, 466, 960

Pentericci L., et al., 2011, ApJ, 743, 132

Pentericci L., et al., 2014, ApJ, 793, 113

Péroux C., Howk J. C., 2020, ARA&A, 58, 363

Pillepich A., et al., 2018a, MNRAS, 473, 4077

Pillepich A., et al., 2018b, MNRAS, 475, 648

Pillepich A., et al., 2019, MNRAS, 490, 3196

Planck Collaboration et al., 2016, A&A, 594, A13

Planck Collaboration et al., 2020, A&A, 641, A6

Puchwein E., Haardt F., Haehnelt M. G., Madau P., 2019, MNRAS, 485, 47

Rahmati A., Schaye J., 2018, MNRAS, 478, 5123

Rahmati A., Pawlik A. H., Raičevìc M., Schaye J., 2013, MNRAS, 430, 2427

Robertson B. E., et al., 2013, ApJ, 768, 71

Rosdahl J., et al., 2018, MNRAS, 479, 994

Schenker M. A., Ellis R. S., Konidaris N. P., Stark D. P., 2014, ApJ, 795, 20

Schroeder J., Mesinger A., Haiman Z., 2013, MNRAS, 428, 3058

Seiler J., Hutter A., Sinha M., Croton D., 2019, MNRAS, 487, 5739

Shapiro P. R., Giroux M. L., 1987, ApJ, 321, L107

Smith A., Safranek-Shrader C., Bromm V., Milosavljević M., 2015, MNRAS, 449, 4336

Smith A., Ma X., Bromm V., Finkelstein S. L., Hopkins P. F., Faucher-Giguère C.-A., Kereš D., 2019, MNRAS, 484, 39

Smith A., Kannan R., Garaldi E., Vogelsberger M., Pakmor R., Springel V., Hernquist L., 2021, arXiv e-prints, p. arXiv:2110.02966

Sobacchi E., Mesinger A., 2015, MNRAS, 453, 1843

Springel V., 2010, MNRAS, 401, 791

Springel V., Hernquist L., 2003, MNRAS, 339, 289

Springel V., et al., 2018, MNRAS, 475, 676

Springel V., Pakmor R., Zier O., Reinecke M., 2021, MNRAS,

Stanway E. R., Eldridge J. J., 2018, MNRAS, 479, 75

Tacchella S., Bose S., Conroy C., Eisenstein D. J., Johnson B. D., 2018, ApJ, 868, 92

Tepper-García T., 2006, MNRAS, 369, 2025

Tilvi V., et al., 2014, ApJ, 794, 5

Totani T., Kawai N., Kosugi G., Aoki K., Yamada T., Iye M., Ohta K., Hattori T., 2006, PASJ, 58, 485

Trac H., Cen R., Loeb A., 2008, ApJ, 689, L81

Trebitsch M., et al., 2020a, arXiv e-prints, p. arXiv:2002.04045

Trebitsch M., Volonteri M., Dubois Y., 2020b, MNRAS, 494, 3453

Vogelsberger M., et al., 2014a, MNRAS, 444, 1518

Vogelsberger M., et al., 2014b, Nature, 509, 177

Vogelsberger M., Marinacci F., Torrey P., Puchwein E., 2020a, Nature Reviews Physics, 2, 42

Vogelsberger M., et al., 2020b, MNRAS, 492, 5167

Walt S. v. d., Colbert S. C., Varoquaux G., 2011, Computing in Science & Engineering, 13, 22

Walther M., Oñorbe J., Hennawi J. F., Lukić Z., 2019, ApJ, 872, 13

Wang F., et al., 2020, ApJ, 896, 23

Weinberger R., et al., 2017, MNRAS, 465, 3291

Weinberger R., Springel V., Pakmor R., 2020, ApJS, 248, 32

Wiersma R. P. C., Schaye J., Smith B. D., 2009, MNRAS, 393, 99

Wise J. H., Cen R., 2009, The Astrophysical Journal, 693, 984

Worseck G., et al., 2014, MNRAS, 445, 1745

Wu X., Kannan R., Marinacci F., Vogelsberger M., Hernquist L., 2019, MNRAS, 488, 419

Wyithe J. S. B., Bolton J. S., 2011, MNRAS, 412, 1926

Xu H., Wise J. H., Norman M. L., Ahn K., O'Shea B. W., 2016, ApJ, 833, 84

Yang J., et al., 2020, ApJ, 904, 26

Zhu H., Avestruz C., Gnedin N. Y., 2020, ApJ, 899, 137

de Belsunce R., Gratton S., Coulton W., Efstathiou G., 2021, arXiv e-prints, p. arXiv:2103.14378

van der Velden E., 2020, The Journal of Open Source Software, 5, 2004

APPENDIX A: COMPARISON WITH THE ILLUSTRISTNG MODEL

We provide in the following a brief comparison between the THESAN model and the IllustrisTNG one, leaving a more thorough investigation to a future work. We perform such comparison by contrasting the results from the THESAN-2 and THESAN-TNG-2 runs, that share the

Table A1. Same as Table 1, but showing the two additional runs used in this Appendix.

Name	L_{box} [cMpc]	$N_{\mathrm{particles}}$	m_{DM} [M$_\odot$]	m_{gas} [M$_\odot$]	ϵ [ckpc]	$r_{\mathrm{cell}}^{\mathrm{min}}$ [pc]	z_{end}	f_{esc}	Description
THESAN-SMALL	23.9	2×525^3	3.12×10^6	5.82×10^5	2.2	~ 10	5.5	0.37	fiducial model (RMHD + TNG + dust)
THESAN-SMALL-TNG	23.9	2×525^3	3.12×10^6	5.82×10^5	2.2	~ 10	5.5	-	MHD + TNG model (original TNG)
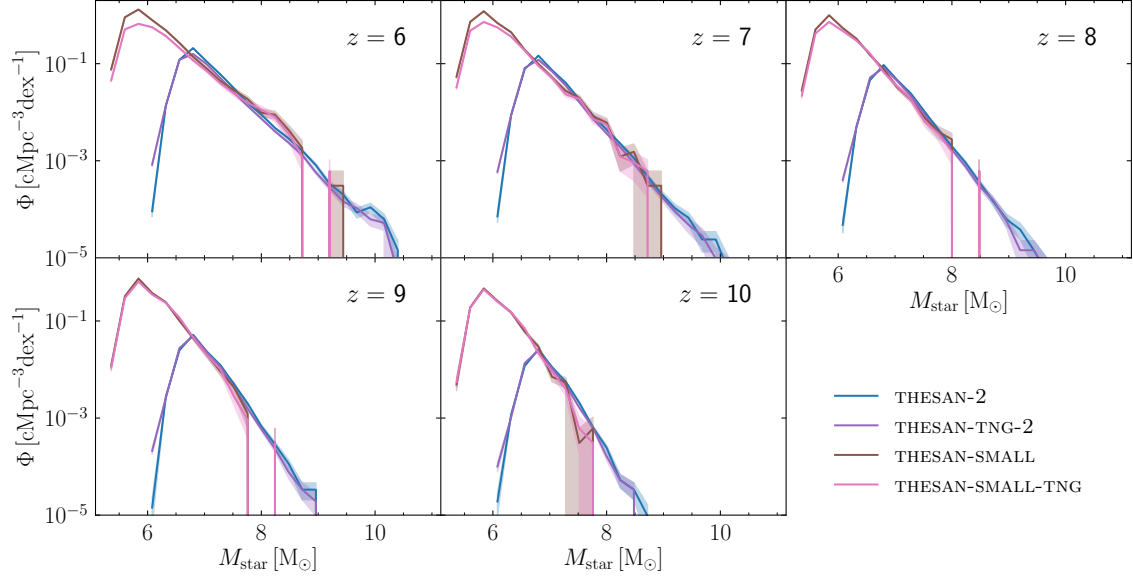
Figure A1. Galaxy stellar mass function for the THESAN-2 and THESAN-TNG-2 runs at different redshifts. The two curves differ only at low stellar masses and only towards the end of reionization.
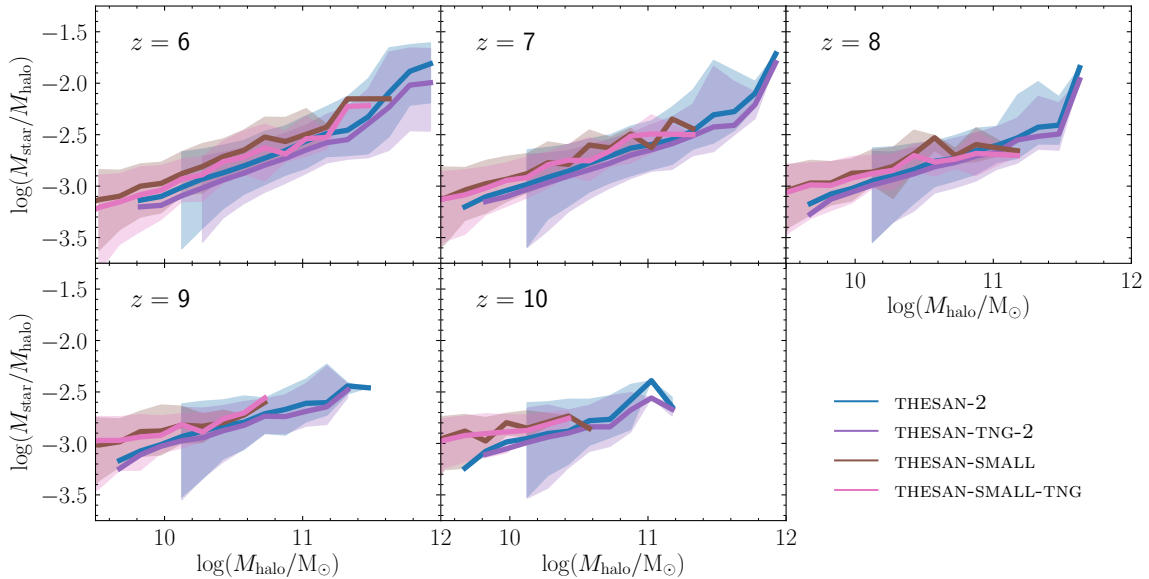
Figure A2. Stellar-to-halo-mass relation for the THESAN-2 and THESAN-TNG-2 runs at different redshifts. The two curves show a systematic offset only at $z \lesssim 8$.
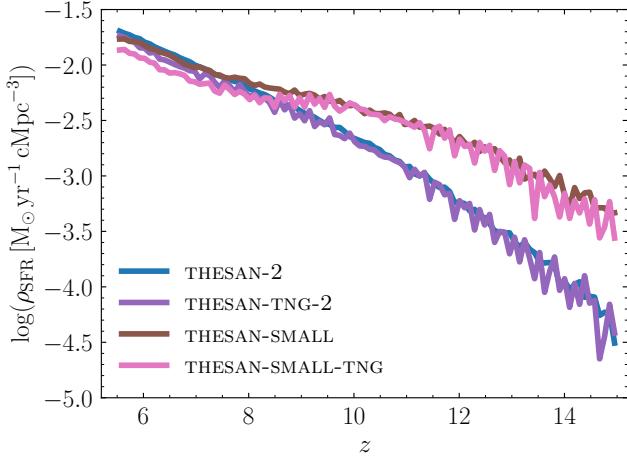
Figure A3. Star formation rate density evolution for the THESAN-2 and THESAN-TNG-2 runs. The two curves differ only at $z \lesssim 8$.

same initial conditions, allowing us to ascribe their differences unequivocally to the different galaxy formation models. The resolution of these runs, however, is lower than our fiducial run, potentially hiding differences in the lowest-mass haloes of THESAN-1. For this reason, we have run two additional simulations covering a smaller volume but reaching the same mass resolution as our flagship run. We name THESAN-SMALL the one employing the THESAN model, and THESAN-SMALL-TNG the one using the IllustrisTNG model. They share the same initial conditions, and their properties are reported in Table A1.

In order to confirm this interpretation, we show in Fig. A2 the stellar-to-halo mass relation. This graph clearly shows that until $z = 7$ the stellar content of haloes is similar in the two models, while at later times the ratio $M_{\rm star}/M_{\rm halo}$ is comparatively (slightly) enhanced in the THESAN model.

In Fig. A1 we show the galaxy stellar mass function in THESAN-2, THESAN-TNG-2, THESAN-SMALL and THESAN-SMALL-TNG. Each pair of curves shows are very similar at all but the smallest stellar masses. At $M_{\rm star} \lesssim 10^6\,{\rm M_\odot}$, the two models show different numbers of galaxies towards the end of reionization (i.e. $z \lesssim 7$). In particular, the THESAN model seems to produce slightly more galaxies than the IllustrisTNG one. Employing a self-consistent radiation transport scheme ensures that haloes are exposed to the correct (spatially-varying) radiation field, which in the case of these small haloes is expected to be mostly generated from external sources. On the contrary, the spatially-uniform optically-thin UVBG employed by IllustrisTNG injects the same radiation in all haloes. Additionally, the latter is developed to complete reionization by $z \sim 6$, while in THESAN this process completes later. Hence, at any given redshift, the average UVBG in THESAN-2 and THESAN-SMALL is lower than in THESAN-TNG-2 and THESAN-SMALL-TNG, reducing the suppression of star formation in small haloes in the former.

Finally, we show in Fig. A3 the evolution of the star-formation-rate density $\rho_{\rm SFR}$, which is indistinguishable between each pair of runs until $z \sim 8$, and subsequently slightly (comparatively) suppressed in the IllustrisTNG model. The difference across the two run pairs is due to their difference resolution, which allows the THESAN-SMALL and THESAN-SMALL-TNG simulations to resolve smaller haloes. These haloes are the dominant source of star formation at very high redshift, explaining the boosted $\rho_{\rm SFR}$ at $z \gtrsim 9$.