IOP Publishing | Institute of Physics and Engineering in Medicine Physics in Medicine & Biology

Phys. Med. Biol. **61** (2016) 7725–7743 doi:10.1088/0031-9155/61/21/7725

Accelerated prompt gamma estimation for clinical proton therapy simulations

Brent F B Huisman[1,2], J M Létang[1], É Testa[2] and D Sarrut[1]

[1] CREATIS, Université de Lyon, CNRS UMR5220, INSERM U1206, INSA-Lyon, Université Lyon 1, Centre Léon Bérard, Lyon, France
[2] IPNL, Université de Lyon, CNRS/IN2P3 UMR5822, Université Lyon 1, Lyon, France

E-mail: brent.huisman@creatis.insa-lyon.fr

Received 11 February 2016, revised 10 August 2016
Accepted for publication 18 August 2016
Published 14 October 2016

Abstract

There is interest in the particle therapy community in using prompt gammas (PGs), a natural byproduct of particle treatment, for range verification and eventually dose control. However, PG production is a rare process and therefore estimation of PGs exiting a patient during a proton treatment plan executed by a Monte Carlo (MC) simulation converges slowly. Recently, different approaches to accelerating the estimation of PG yield have been presented. Sterpin *et al* (2015 *Phys. Med. Biol.* **60** 4915–46) described a fast analytic method, which is still sensitive to heterogeneities. El Kanawati *et al* (2015 *Phys. Med. Biol.* **60** 8067–86) described a variance reduction method (pgTLE) that accelerates the PG estimation by precomputing PG production probabilities as a function of energy and target materials, but has as a drawback that the proposed method is limited to analytical phantoms.

We present a two-stage variance reduction method, named voxelized pgTLE (vpgTLE), that extends pgTLE to voxelized volumes. As a preliminary step, PG production probabilities are precomputed once and stored in a database. In stage 1, we simulate the interactions between the treatment plan and the patient CT with low statistic MC to obtain the spatial and spectral distribution of the PGs. As primary particles are propagated throughout the patient CT, the PG yields are computed in each voxel from the initial database, as a function of the current energy of the primary, the material in the voxel and the step length. The result is a voxelized image of PG yield, normalized to a single primary. The second stage uses this intermediate PG image as a source to generate and propagate the number of PGs throughout the rest of the scene geometry, e.g. into a detection device, corresponding to the number of primaries desired.

We achieved a gain of around 10^3 for both a geometrical heterogeneous phantom and a complete patient CT treatment plan with respect to analog MC,

at a convergence level of 2% relative uncertainty in the 90% yield region. The method agrees with reference analog MC simulations to within 10^{-4}, with negligible bias. Gains per voxel range from 10^2 to 10^4.

The presented generic PG yield estimator is drop-in usable with any geometry and beam configuration. We showed a gain of three orders of magnitude compared to analog MC. With a large number of voxels and materials, memory consumption may be a concern and we discuss the consequences and possible tradeoffs. The method is available as part of Gate 7.2.

Keywords: prompt gamma, proton therapy, Monte Carlo, variance reduction techniques

(Some figures may appear in colour only in the online journal)

1. Introduction

The well defined range of particles in matter is the main reason they are used in cancer treatment today. Unfortunately we are not able to take full advantage of this property, because of uncertainties in patient positioning, uncertainties in the proton range due to unknown displacements or deformations of organs, ill defined lung, bowel or bladder filling, and the inherent uncertainty in the Hounsfield unit to particle stopping power conversion. Often, medical practice is to plan conservatively, namely adding margins around the tumor, greatly reducing the potential benefits of particle treatment (Knopf and Lomax 2013). Some form of *in vivo*, online monitoring is generally considered to be a way out of this predicament. Online monitoring would make measurements of uncertainties such as mentioned above possible, and thereby permit more precise planning, which could take maximum advantage of the steep Bragg peak (BP) falloff and reduce damage to tissues surrounding the tumor.

One way to perform monitoring is to use prompt gammas (PGs), a natural byproduct in particle treatments. PGs offer the potential of monitoring treatment at the spot level (Smeets *et al* 2012, Roellinghoff *et al* 2014) and are therefore of prime interest (Moteabbed *et al* 2011, Gueth *et al* 2013, Golnik *et al* 2014, Janssen *et al* 2014). These particles are produced in the inelastic nuclear collisions between the incident particle and the medium (tissue) it is traveling through, and they correlate very well with dose deposition. This has been experimentally demonstrated for protons by Min *et al* (2006) and for carbon by Testa *et al* (2008). The latter also showed that adding time of flight (ToF) measurement enables discrimination between neutron-induced gammas and PGs, greatly cleaning up the signal, which was confirmed by Biegun *et al* (2012), Lopes *et al* (2015), Pinto *et al* (2015), Roellinghoff *et al* (2014), Smeets *et al* (2012) and Verburg and Seco (2014). The authors conclude that online range monitoring is possible, once a detector with sufficiently large solid angle can be realized. In Fall 2015, at OncoRay in Dresden, Germany, a knife-edge PG camera (Perali *et al* 2014, Richter *et al* 2016) was put into clinical operation. In addition to geometrical collimation, work is being done on designs that exploit Compton scattering to record the PG signal (Roellinghoff *et al* 2011, Kurosawa *et al* 2012).

Another approach is to use the fact that in certain nuclear reactions PGs with specific energies are produced (Verburg *et al* 2013, Verburg and Seco 2014). The composition of the produced PG spectrum is dependent on material and proton energy. Certain PG energies tend to be produced close to the BP; that is to say, the associated cross sections are largest when the primary is near the end of its range. Analysis of these spectral lines at given positions may provide sufficient information to deduce both the material composition and BP position. Verburg

et al (2013) exploits this fact to reconstruct the spatial location of the BP falloff position with a single collimated detector placed at the end of the primary range. Then there is an attempt (Golnik *et al* 2014) to use the travel time of the primary to reconstruct and control the PG range. Using a camera with high timing resolution and a short, pulsed beam, the ToF between the proton leaving the beam nozzle and the PG arriving in the camera can be recorded, and by putting the camera close to the nozzle looking back at the patient we know that a larger ToF corresponds to more distal interactions, and a smaller ToF to proximal interactions.

Imaging paradigms such as PG detection are validated against experiments, and often also with Monte Carlo (MC) simulations (Moteabbed *et al* 2011, Gueth *et al* 2013, Robert *et al* 2013, Golnik *et al* 2014, Janssen *et al* 2014). Conventional MC methods propagate particles according to a set of physical processes through materials. The tracking of a particle is broken down into steps, where at each point a weighted list of all possible next steps is built and one option is selected by a random number. For rarely occurring processes, convergence to the model of the truth to within acceptable statistical error can be slow. PG emission in particle therapy, when viewed on a voxel-by-voxel basis, is also a rare and slowly converging process (Gueth *et al* 2013, Pinto *et al* 2015, Sterpin *et al* 2015). This has important implications: first for detector designers, second for those who make simulations, and third for those interested in comparing the two online in, say, clinical conditions. In the first case, detectors are optimized to minimize signal loss (see figure 1) and advanced reconstruction can be employed to maximize the use of information in the signal. Gueth *et al* (2013) has demonstrated a method that works around the low PG yield by using machine learning to correlate predefined patient translations (setup errors) to PG output signals, which reduces the time to produce an estimated translation based on the detected PGs. Since convergence to the model of the truth requires long simulation runtimes, we may compensate with variance reduction methods or cutoffs in the time, space, or spectral domains (e.g. fixed runtime, larger voxel size, larger spectral bins). One such variance reduction method is MC splitting, where the moment a rare process occurs not one, but multiple possible futures for that process are computed. The weighted total of these futures is then stored, and thus the convergence accelerated.

Recently, a variance reduction method (pgTLE) was described (El Kanawati *et al* 2015). The PG estimation is accelerated by precomputing PG production probabilities as a function of energy and target materials, with the drawback that it only works for analytical phantoms. Here, we present a two-stage method, voxelized pgTLE (vpgTLE), that extends pgTLE to voxelized volumes, making it useful for clinical simulations. Next we describe the method and give some simulation results in various configurations. At this time, only protons as primaries have been considered and validated. We finish with a discussion of background noise estimation, other variance reduction methods and points for improvement.

2. Materials and methods

2.1. A voxelized pgTLE

A full *voxelized prompt-gamma track length estimator* (vpgTLE) simulation is broken up into two stages (figure 2). The process presumes the existence of a prepared database (PGdb), which is an estimate of the effective linear PG production coefficient modulo the density, per element ($\frac{\Gamma_Z}{\rho_Z}$, see equation (1)), per PG energy bin per primary energy bin. The PGdb is computed once for a list of common elements, and can then be reused. In stage 1, a PGyd image is created, specific to a particular phantom (or CT image) and primary source (e.g. a treatment plan, a single spot). This PGyd image stores, per voxel per PG energy bin, the yield

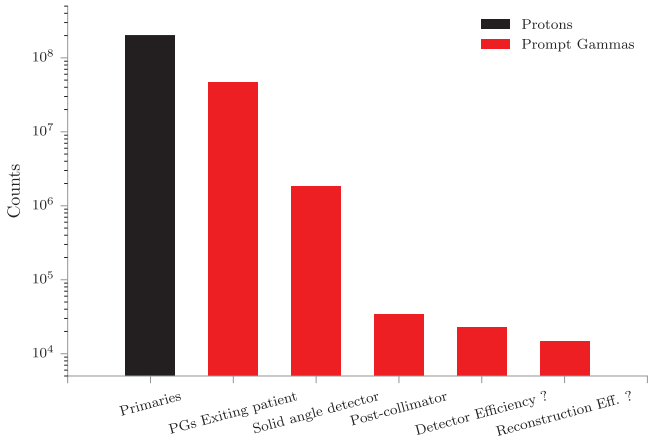

Figure 1. Approximate progressive loss of PG signal per spot. The particles were recorded in a simulation with a clinical head and neck CT and associated clinical plan, with a 30×30 cm^2 collimator with 43% fill factor, at 40 cm from the patient, with estimated efficiencies (66%) for detection and reconstruction. Note this is not the configuration investigated in the rest of this paper.

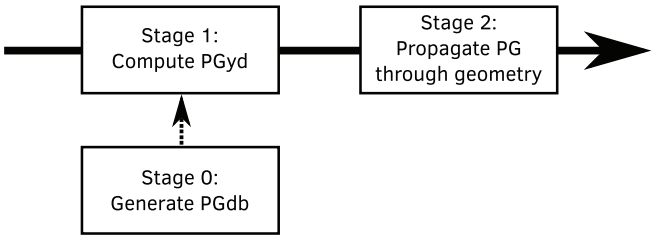

Figure 2. Flow chart of the vpgTLE method. Stage 0 is an initial PG database (PGdb) and is computed once. Every subsequent simulation is broken up into stages 1 and 2. Stage 1 generates a CT-specific PG yield distribution (PGyd) using a limited number of primaries. In stage 2 the PGyd is used to generate and propagate a representative set of PGs throughout the geometry.

per primary. Stage 2 uses the PGyd and the assumption of isotropic PG emittance to generate and propagate the PGs throughout the rest of the geometry that the user has defined (e.g. the PG detector).

2.1.1. Stage 0: PG database. First, as explained in El Kanawati *et al* (2015), we precalculate a database of PG production probability per PG energy, as a function of element and proton energy. Equation (1) describes the computed quantity: the PG spectrum $\boldsymbol{\Gamma}$ as a function of proton energy E for a set of elements Z. We take the ratio of the number of produced gammas $\boldsymbol{N}_\gamma$ over the number of inelastic collisions N_{inel}, scaled by the linear element attenuation coefficient κ_{inel}, related to the proton inelastic nuclear process. We normalize with the density ρ of element Z. Bold symbols represent vectors as function of PG energy. We compute the PGdb for protons up to 200 MeV in 250 bins.

$$\frac{\boldsymbol{\Gamma}_Z(E)}{\rho_Z} = \frac{\boldsymbol{N}_\gamma(Z,E)}{N_{\text{inel}}(Z,E)} \frac{\kappa_{\text{inel}}(Z,E)}{\rho_Z}. \tag{1}$$

Currently, PG emission models and cross sections implemented in various MC codes (Geant4, FLUKA, MCNPX) are still evolving, as differences have been observed between experimental PG data and simulations (Pinto *et al* 2015) and between MC codes (Verburg *et al* 2012, Robert *et al* 2013, Pinto *et al* 2016), and research is ongoing to improve accuracy. Hence, if cross sections or implementations of models are updated, the database must be recomputed. In particular, when comparing between simulations, such as in this study between vpgTLE and a reference analog MC, the same physics list must be used.

2.1.2. Stage 1: compute phantom-specific PG yield distribution. By performing a low statistic MC, an estimate of the proton track lengths per voxel per proton energy bin is obtained, hence 'track length estimation'. Note that 'track' corresponds to 'step' in Geant4 terminology. Together with the effective linear PG production coefficient, the PGyd image is obtained.

Before the simulation starts, we query the CT image for a list of present materials and build $\mathbf{\Gamma}_m$ for each material m encountered. We compute the PGdb in terms of elements, rather than materials, to give the user the option to add new materials or modify the composition of existing ones without recomputing the PGdb. In equation (2), we sum $\mathbf{\Gamma}_Z$ over all k constituent elements Z in material m_v, weighted by density fraction ω_k of element Z in m_v, and scaled by density ρ_{m_v}.

$$\mathbf{\Gamma}_m(E) = \rho_{m_v} \sum_{k=1}^{k_{m_v}} \omega_k \frac{\mathbf{\Gamma}_{Z_k}(E)}{\rho_{Z_k}}. \tag{2}$$

Then, a limited number of particles are propagated from the source of the treatment plan into the target, typically a patient CT image. We define a 4D scoring matrix in the CT (the PGyd), which may have a different size and pixel spacing from the CT. Protons are propagated with an unmodified analog MC tracking engine. Per step, per voxel v in the PGyd, alongside executing the unmodified analog MC processes, we compute and add the product of the step length L_g and $\mathbf{\Gamma}_{m_v}$ (linearly dependent on proton inelastic cross section, see equation (1)), with m_v the material at voxel v and g the proton energy bin, according to equation (3). Put into words, we compute the PG yield probability energy spectrum at every step, and add it to the yield of voxel v of the current step.

$$\widehat{\boldsymbol{S}}_i(v) = \mathbf{\Gamma}_{m_v}(E_g) L_g(E_g, v). \tag{3}$$

At the end of the simulation, we have accumulated the yield spectra per voxel v. The computed output is weighted by the number of primaries to obtain the final PG production probabilities per voxel per PG energy bin. The PGyd written to disk is therefore *per primary*, and the sum of all the probabilities is the probability that a single primary particle will emit a PG.

In order to obtain the variance in this paper, we opted for the classic batch technique. Although the authors of El Kanawati *et al* (2015) provide an analytical derivation for the variance, they assume no correlation between proton energy and track lengths. We observed that this assumption does not seem to hold, in that the result is quite different from the variance obtained with the batch technique. Derivations assuming partial and full correlation are possible, but we felt that the short runtimes of the vpgTLE method, coupled with the simplicity of the batch technique, and the innate correctness of the variance obtained in such a way, were the best choice for understanding and presenting this method. Note that this assumes the initial database computed in stage 0 has converged sufficiently and does not contribute significantly to the variance.

2.1.3. Stage 2: propagate PG through other geometry (detectors). The PGyd computed in stage 1 is used as a source of PG emission probability per primary particle. All that is left to complete the simulation (e.g. record PGs in a detector) is to produce as many PGs weighted

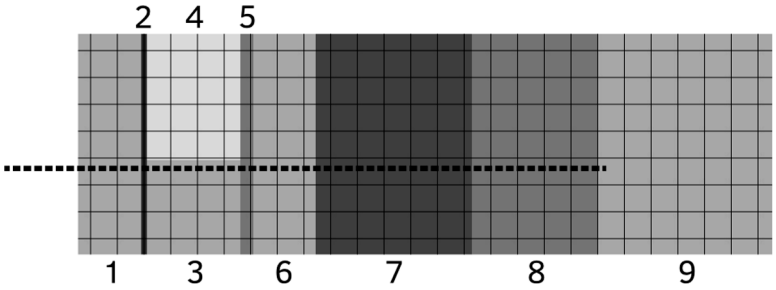

Figure 3. Top-down view of the Parodi phantom (Parodi *et al* 2005), where the shading represents the material densities. Parts 1, 3, 6, 9 are PE; 7 is PMMA; 2 is bone; 5, 8 are muscle; 4 is lung. The beam is illustrated with the dotted line coming in from the left. A voxelized version of this image is created, with a 2^3 mm^3 voxel size.

by the source image as necessary. If, say, the user is interested in the PG signal of a 2 Gy fraction, and a 2 Gy fraction is composed of 10^{11} protons, the PGyd can be requested to give the expected output for that number of protons. Each PG is created and then propagated through the patient and into the detector, according to regular analog MC mechanisms. However, depending on the number of PGs users require in their detectors, a lower number of PGs may be requested and scaled up, and consequently runtime is lowered.

An important consideration is that we currently assume that PGs are spatially emitted isotropically. Geant4 also adheres to this assumption. This conveniently relieves us from recording any spatial information. However, there exists evidence to the contrary (Sheldon and Van Patter 1966, Verburg *et al* 2012). Once any possible anisotropy is introduced into the MC physics models, it can be added to our code. Recording the anisotropy factor during stage 1 may be an intuitive solution (Henyey and Greenstein 1941).

2.2. Validation procedure

Each simulation is executed both with analog Monte Carlo scoring and with the vpgTLE method. The analog MC serves as reference. To obtain an estimate of the statistical uncertainty, we employ the batch technique and run each type of simulation 10 times. When studying the bias and the relative uncertainty within selected subregions of the phantoms, σ is computed on the projection considered. That is to say, σ represents the standard deviation on the mean yield over the 10 simulations. For the study on efficiency and convergence of relative uncertainty, the variance is computed per voxel. Taking the median of (a subregion of) this 4D 'image of variance' (i.e. the median of the variance) provides a stronger test. We take the median rather than the mean because the variance distribution tends to a log-normal distribution. For skewed distributions such as this, the median is a better measure of central tendency.

2.2.1. Test cases. Two test cases are presented. The purpose of test case 1 is to verify that the transition to voxels has been done correctly. The phantom proposed by Parodi *et al* (2005) and used by El Kanawati *et al* (2015) is converted into a voxelized representation; see figure 3. In test case 2 (figure 4), a clinical head and neck image with corresponding proton therapy treatment plan is examined and is intended as a demonstration of the possibilities of the vpgTLE method. Precise beam properties may be found in table 1. Compared to El Kanawati *et al* (2015), the number of analog primaries used for the reference is increased from 10^7 to 10^9. This is required in order to obtain a sufficiently noiseless figure for the spatial projection along the beam axis. For the vpgTLE simulations, four simulations are executed with 10^3, 10^4, 10^5

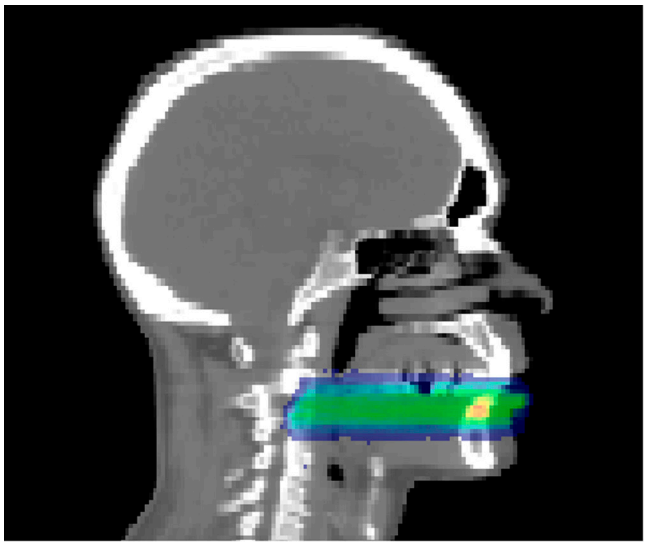

Figure 4. Sagittal view of the patient CT image, illustrated with the PG yield caused by the associated treatment plan. A beam from an original treatment plan has been rotated to align with the image axes, both to make projections easier and to increase the heterogeneity of the materials and thereby increase the challenge to vpgTLE. The applied radiation is the distal layer of one of the beams of the original plan so that the distal falloff is better defined in contrast to the spread out Bragg peak (SOBP). The voxel size is again 2^3 mm^3.

and 10^6 primaries respectively. Next, we define certain windows of interest. Knowing that most PG detectors do not measure outside the 1–8 MeV energy range (Testa *et al* 2013), or even narrower (Smeets *et al* 2012), we limit our analysis to this energy window. In addition, PG yield outside the spatial region that represents 90% of the total yield in the reference simulation is discarded. For all analyses these two cuts are applied.

2.2.2. Bias. To establish the presence of bias, as a function of spatial or spectral dimensions, the relative difference with respect to the reference is investigated. In addition, certain subregions are studied separately for test case 1, because material-based regions can easily be isolated.

2.2.3. Efficiency, gain and convergence. An important quantity that characterizes a variance reduction method is the efficiency ϵ_k, which is computed by considering the time t required to reach a variance σ_k^2 per voxel k; see equation (4). Comparing the ratio of efficiencies of two methods gives a quantified gain G (equation (5), where TLE and A refer to vpgTLE and analog MC respectively). Using a measure of centrality (e.g. mean, median) on the gains per voxel G_k, a global measure for the efficacy of vpgTLE is obtained. A histogram of the gains within an image is presented to give an idea of the worst and best case performance of vpgTLE.

$$\epsilon_{\mathrm{TLE},k} = \frac{1}{t \times \sigma^2_{\mathrm{TLE},k}} \tag{4}$$

$$G = \frac{\epsilon_{\mathrm{TLE},k}}{\epsilon_{A,k}}. \tag{5}$$

Table 1. Summary of vpgTLE analysis parameters.

	Test case 1	Test case 2
Beam	160 MeV, disc shaped cross section, Gaussian spatial and angular distr. with σ of 3.5 mm, 2 mrad respectively	Distal layer (133.08 MeV) of clinical treatment plan (7 spots)
Phantom description	Parodi *et al* (2005)	Clinical head and neck CT
Phantom voxel size	2 mm	
PGyd voxel size	2 mm	1, 2, 5 mm
PGdb primaries/element	10^9	
PGdb max. proton energy	200 MeV	
Number of PG bins	250	
Number of proton bins	250	
vpgTLE primary sets	10^3–10^6	
Analog primary sets	10^6–10^9	
Reference	Analog primary set 10^9	
Studied projections	Spectral and along beam	
Extra studies	Central voxel line yield	Influence of PGyd voxel size
Spatial window	90% yield region	
Spectral window	1–8 MeV	
Variance computation	10 batches per primary set	
Step size	1 mm	

The proposed vpgTLE is a variance reduction method: it reduces the variance for any given PG simulation with respect to an analog simulation. As a simulation runs, the output converges, which is to say its variance reduces. A common measure to ensure sufficient convergence is to require that the uncertainty σ associated with the quantity of interest is not more than 2% of the signal. A variance reduction technique such as vpgTLE translates into reaching the threshold faster, and therefore a gain with respect to analog MC. We compute the gain therefore both by taking the ratio of the runtime of the two methods at the 2% convergence level and by computing the gain (equation (5)). With the same data, we estimate the total runtime required to produce a sufficiently converged image.

2.2.4. Influence of voxel size. An essential property of vpgTLE is the fact it records PG production probabilities all along the primary's path, instead of waiting for the MC engine to produce a PG. This means that vpgTLE requires far fewer propagated primaries to touch all voxels. This effect could be magnified further when smaller voxels are used. Recent developments (Marcatili *et al* 2014) in super-resolution or multi-scale CT models involve smaller voxels in order to increase simulation accuracy for smaller or thinner organs such as the rectum, bladder or spine. We demonstrate the increased benefits of vpgTLE on test case 2.

2.2.5. Hardware, software, parameters. Gate 7.1 (Jan *et al* 2004, Sarrut *et al* 2014) with Geant 4.10.01 and the QGSP_BIC_HP_EMY physics list, commonly used for PG studies, are used in this analysis.

In Gate, scorers are defined as actors, which can be attached to volumes, all defined through the Gate macro language. The vpgTLE method was implemented as an actor, that keeps a PG spectrum for each voxel, the number and volume of which can also be controlled through the macro language. A helper actor, that outputs the analog MC in a similar manner as the PGyd, was implemented to facilitate analysis. The output was validated against the GatePhaseSpaceActor, a thoroughly used and tested part of Gate. We used a gamma production cut of 3 keV in order to remove a high number of photons that cannot exit the phantom or patient geometry.

Timing information is obtained with the aid of the GateSimulationStatisticActor, executed on an Intel Core i7-3740QM CPU @ 2.70 GHz, SpeedStep off, whilst using a single core. In summary, table 1 lists the main parameters used for the simulations.

3. Results

3.1. Test case 1

First we verified that vpgTLE yields are identical to the results produced with pgTLE, shown in El Kanawati *et al* (2015). Then, we compared our method to the analog reference. Figure 5 depicts the yield on the first row, as a function of the depth (left column) and the PG energy (right column). The relative difference of the PG yield is shown, integrated either over the entire coronal plane (second row) or at the voxel line on the beam path (third row). Row 1, a plot of the yield, shows a perfect overlap of vpgTLE with respect to the reference. We must look to the relative differences of the various vpgTLE outputs with respect to the reference in row 2 to observe any differences. The shaded areas represent 2σ error bands.

With 10^3 primary particles, the mean is noisiest, as expected. An overestimate beyond 170 mm is visible, which is about the location of the Bragg peak. The average relative difference over the depth is 4.0×10^{-4} along the beam, which is a good performance, but due to the relative difference from the reference exceeding 1% in the distal region and the very wide error bands we would argue that 10^3 primaries are insufficient for a reliable prediction. The distal systematic shift reduces when using 10^4 or more primaries. Two regions with bias remain: a consistent overestimate of about 0.5% at around 160 mm depth, and then, past the Bragg peak, an erratic mean with wide uncertainty bands. The latter can be explained by nuclear events. Once a proton collides and is absorbed, it can no longer produce PGs. Towards the end, the precise number of remaining protons grows more uncertain, and just 10^3 primaries are not enough for a good estimate of the variance. Increasing the number of primaries reduces the uncertainty and improves the mean yield, but the effect remains. The average relative difference over the depth is of the order of 10^{-4} for all primary sets.

The spectral column on the right demonstrates that vpgTLE is close to the analog reference over the whole spectrum, with a small fluctuation at the high end of the spectrum. The pattern present for all primary sets must be due to the PGdb, and is the only bias we observe. The average relative difference varies between -1.3×10^{-4} with 10^3 primaries and 6.0×10^{-5} for 10^6 primaries. This supports the hypothesis that we have converged to the bias introduced by the uncertainty of the PGdb. The wide error bands for 10^3 are again visible, with the error bands for 10^4 or more primaries staying within 1% of the mean over the entire range.

The bottom row, the line of voxels centered on the beam path, shows a more erratic behavior. One major difference is the proximal overestimate and distal underestimate with 10^3 primaries. With 10^5 primaries or less, the average relative difference is of the order of 10^{-3} or more, while 10^6 primaries result in 3.6×10^{-4}. The uncertainty is, naturally, larger. The

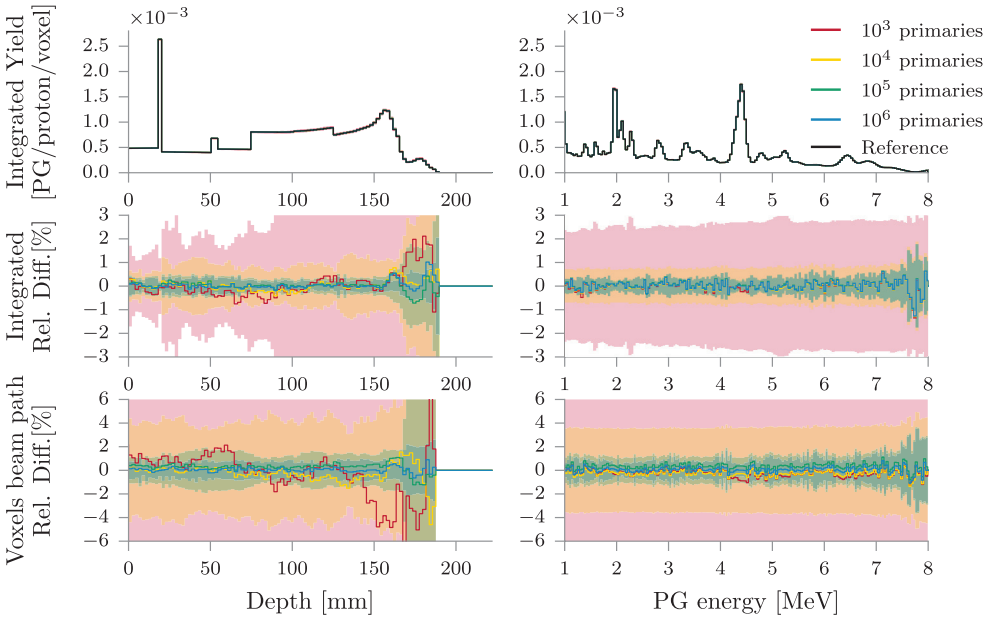

Figure 5. Test case 1: row 1 shows the PG yields and row 2 the relative difference with respect to the reference ($\frac{\text{vpgTLE} - \text{Reference}}{\text{Reference}}$). The yield corresponds to the mean over 10 simulations. For both rows, note that the yield was integrated over all other dimensions. Row 3 shows the relative difference on the line of voxels on the center of the beam path, where we did not integrate over all other dimensions. The left column is a projection along the beam-axis, while the right column shows the spectral bins integrated over all voxels considered. The shaded areas represent 2σ error bands, where σ is the standard deviation over the mean of 10 simulations. Note that the covarying pattern in the relative difference is due to the noise of the analog signal, and does not represent any issue with the vpgTLE implementation.

spectral view is stable and the depth view has an increased variance towards the end of the proton range.

Figure 6 shows side by side the convergence of the median relative uncertainty and a histogram of the gains computed with equation (5). We see that a median convergence to within 2% is reached in about 3 min and about 68 h with vpgTLE and analog MC respectively. At the 2% convergence level, the gain is 1.55×10^3. The histograms on the left show that the gains are stable with respect to the number of vpgTLE primaries. This means that vpgTLE has no systematic problems. We can clearly see the skew of the distributions (note the logarithmic scale on the x-axis). The worst gain is a factor of 6.19×10^1, while the best voxel clocks in at 5.21×10^4, with a median of 1.40×10^3.

3.2. Test case 2

See figure 7 for the yield and relative difference of vpgTLE with respect to the reference. The width of the 2σ-bands has increased with respect to test case 1. Along the beam (left column of the figure) we see that 10^3 primaries produce an erratic line, while 10^5 and higher are close to 0%. Past the distal end, we see significant divergence as in test case 1. While the average bias is of similar magnitude as in test case 1 (10^{-4}, except for 10^3 primaries), on the distal end the bias has not quite dissipated with 10^6 primaries. A likely explanation for the worse

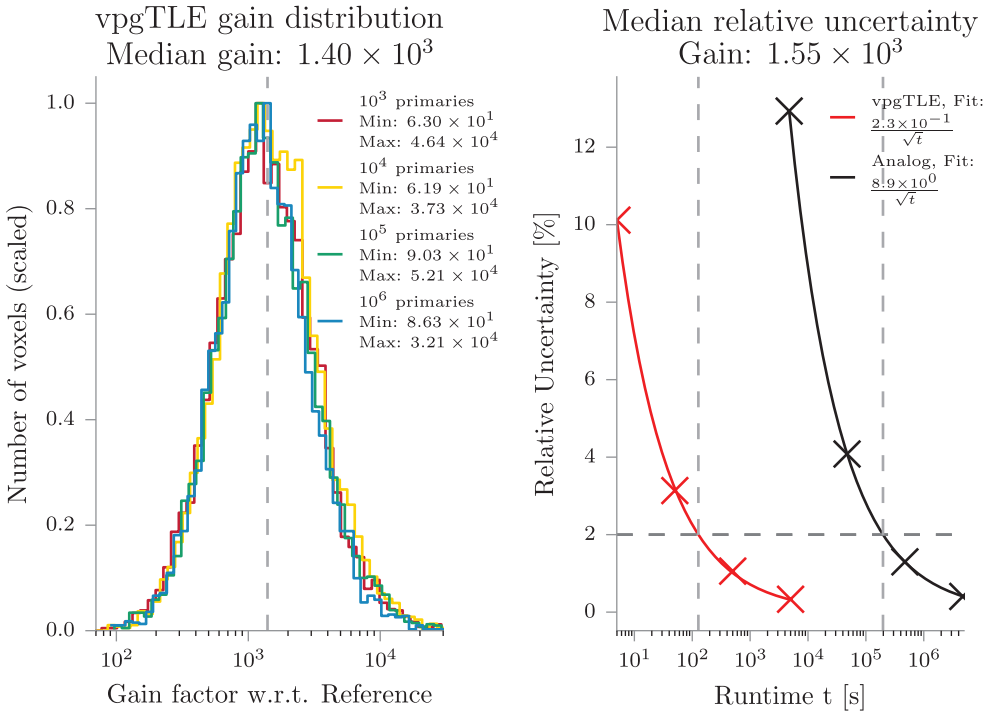

Figure 6. Test case 1. Left, the efficiency computed, per voxel, according to equation (4), for all vpgTLE primary sets with respect to the reference. Right, the median relative uncertainty as a function of runtime t, for both the analog and vpgTLE methods. Each successive point is generated with 10^3–10^6 primaries for vpgTLE, and with 10^6–10^9 primaries for analog MC. As t increases, the relative uncertainty decreases as $\frac{C}{\sqrt{t}}$, where C is a fit factor. To compute the gain, we take the ratio of the runtimes at the 2% level, indicated by the dashed horizontal line, generally considered to be sufficiently converged.

performance is the different beam with respect to test case 1: now we look at the yield caused by a full energy layer composed of seven spots, resulting in the primaries being spread over a larger volume and therefore in lower statistics per voxel. We can again attribute part of the increase in error to the systematic error induced by the PGdb. The spectral view is as stable as in test case 1, which is spread over the same 250 bins, and differs only by wider error bands.

Figure 8 shows the convergence and gain of vpgTLE for test case 2. The gain is slightly less than in test case 1. A sufficiently converged PGyd now requires a little over 4 h on a single core with vpgTLE and about 180 days with analog MC. Excluding the 10^3 primary set because of its outliers, the worst gain is 2.70×10^1 and the best is 8.96×10^4, with a median gain of 9.98×10^2, and a gain of 1.03×10^3 at 2% relative uncertainty.

The last result is the effect of changing the PGyd voxel size on the gain. We have observed in figures 6 and 8 that the gain distribution is independent of the number of primaries. Hence, in order to save time, we conduct this investigation with 10^7 primaries for analog MC and 10^4 for vpgTLE, which are at similar levels of convergence (see figure 8). The gain ratio is computed as the ratio between these sets, per voxel size. Due to memory consumption considerations, the minimum voxel size was set at 1 mm^3, which results in an image of 1666 MB (833 MB on disk).

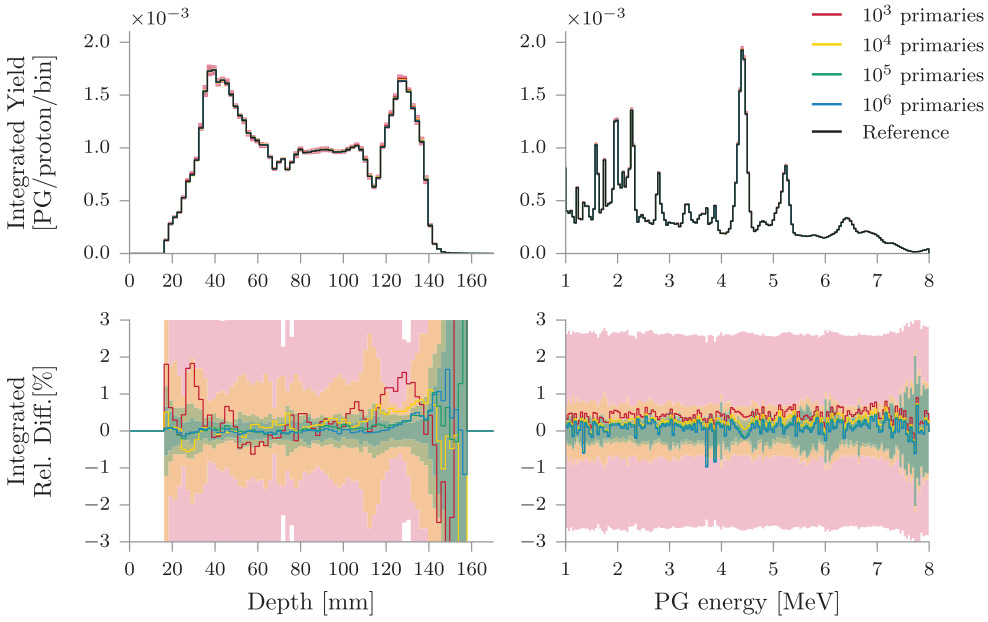

Figure 7. Test case 2. Row 1 shows the PG yields and row 2 the relative difference with respect to the reference ($\frac{\text{vpgTLE} - \text{Reference}}{\text{Reference}}$). For both rows, note that the yield was integrated over all other dimensions. The left-hand column is a projection along the beam axis, while the right-hand column shows the spectral bins integrated over all voxels considered. The shaded areas represent 2σ error bands, where σ is the standard deviation over the mean of 10 simulations.

Figure 9 shows the gain histograms for each voxel size. It can be seen that the advantage of vpgTLE is magnified with respect to analog MC as voxel size decreases. This is in agreement with the hypothesis that vpgTLE needs fewer primaries with respect to analog MC to estimate the PG yield. Users interested in super-resolution output can expect higher gains than reported for our two test cases.

4. Discussion

4.1. Tradeoffs

The current implementation of vpgTLE stores each bin as a double (64 bit) in memory, and converts to float (32 bit) when writing to disk. The memory consumption is therefore linked to the number of PG bins, image and voxel size. By default the vpgTLE actor will copy the size and voxel size of the image it is attached to. As described before, for clinical CT images this will result in on-disk images of tens of gigabytes, and twice that in memory usage during the simulation. In this paper, we therefore shrank the PGyd volume to a region that envelops the PG production sites with some margin (see figure 10). This resulted in an on-disk image size of about 104 MB. With 1 mm^3 voxels, the image size increases eightfold to 833 MB (double at runtime: 1666 MB). A PGyd with the size of the CT data used in test case 2 with 1 mm^3 voxels would blow up to 120 GB on disk, 240 GB in memory at runtime.

Storing the intermediate PGyd is similar to the practice of storing intermediate phase spaces in complex accelerator simulations. A nice side-effect of having two stages is that if

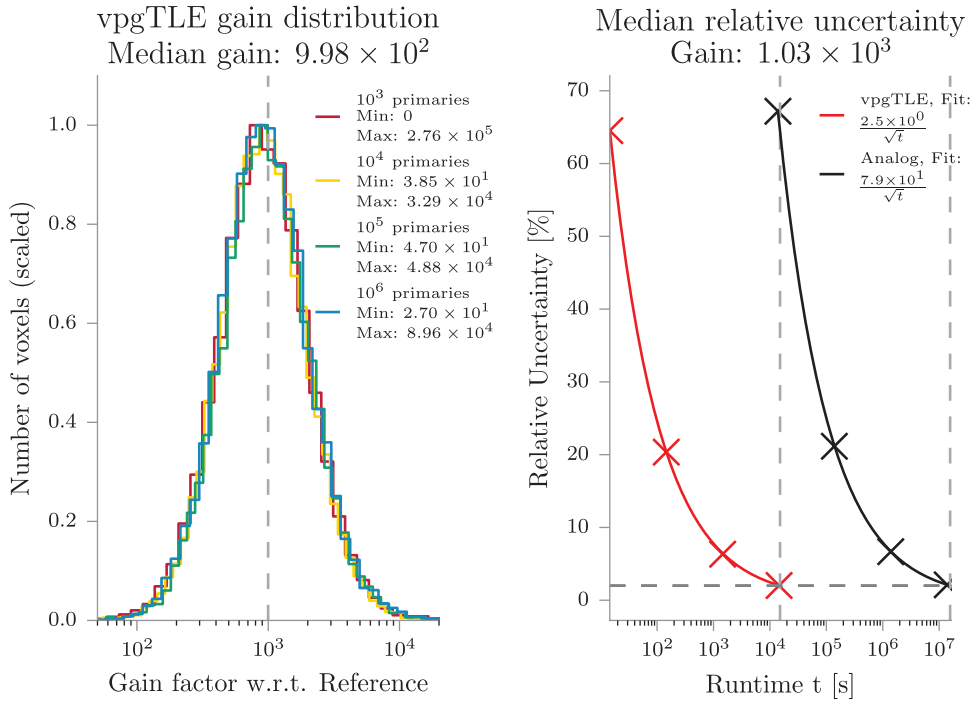

Figure 8. Test case 2. Left, the gain histogram is shown, for all vpgTLE primary sets with respect to the reference. Right, the mean relative uncertainty is plotted as a function of runtime, for both the analog and vpgTLE methods. Each successive point is generated with 10^3–10^6 primaries for vpgTLE, and with 10^6–10^9 primaries for analog MC. We take the ratio of the runtimes at the 2% level to obtain the gain.

the user for instance wants to reposition a detector, or compare different detectors altogether, only the PGs have to be re-propagated.

Before stage 1, a number of important parameters are set: the number of primaries per element in the PGdb, and the minimum, maximum and number of bins for both the primary and PG energy. Naturally, more primaries increase the quality of the estimate, while more bins spanning a wider range improve the precision, but slow down the convergence to an acceptable mean or median uncertainty. Assuming a maximum primary energy of 250 MeV, we need 250 bins for a precision of 1 MeV. The current implementation has linear binning, so assuming that the location of the BP falloff is of interest, where the primary energy is lowest, a 1 MeV bin translates to a proton range in water of about 24 μm, more than enough considering the typical 2^3 mm^3 voxel size. Note that computing the PGdb for more bins requires more particles to ensure proper bin filling. It took approximately 1000 days of CPU-time to compute the PGdb used in this study.

4.2. Comparison with other variance reduction techniques

A conventional approach to variance reduction for rare processes is interaction biasing (IB), where the probability of the interaction of interest is multiplied by factor α, and is compensated for by decreasing the weight of the continuing track (and secondaries) by factor $\frac{1}{\alpha}$. Parameter α is then chosen such that an interaction occurs once per interval of interest (say,

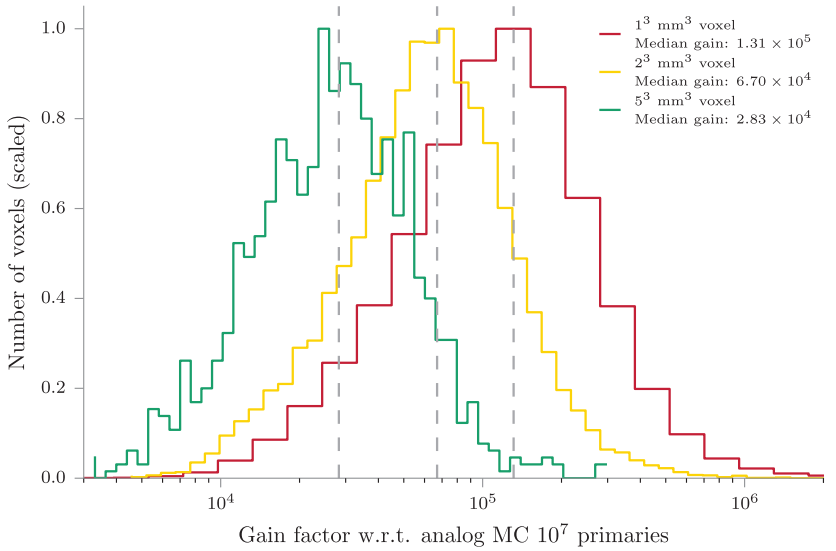

Figure 9. The distribution of gains of a vpgTLE simulation (10^4 primaries) w.r.t. an analog MC (10^7 primaries) is plotted. The distribution, and the medians, shift up as the voxel size decreases.

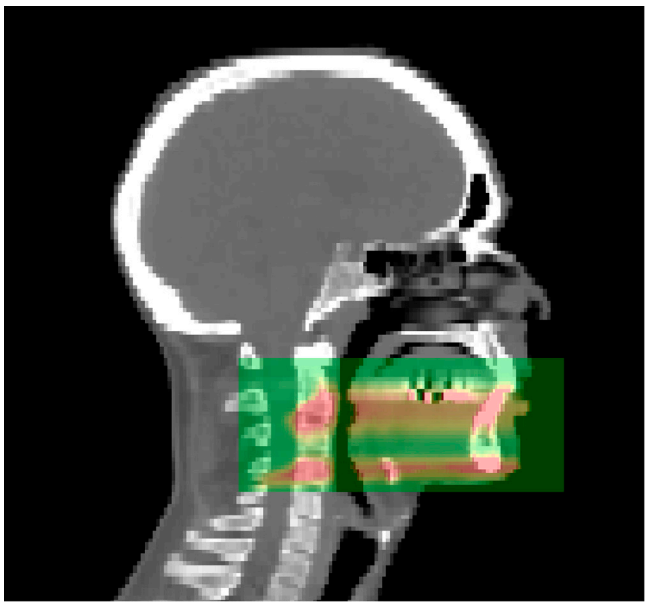

Figure 10. The region of the PGyd was set to a smaller region than the patient CT image, in order to reduce memory consumption. The region of the PGyd is visible as green, with the PG production visible as red bands.

once per voxel). Alternatively, interaction forcing (IF) forces an interaction per interval, and weights any subsequent interaction with the probability that former interaction occurred. As the incident particle may be killed in the process of interest (as is the case for PG production), some implementations (e.g. MCNPX, Geant4, EGSNRC) split the track into a collided and an uncollided version to prevent the loss of statistics in the distal part of the track.

Another standard technique for rare processes is particle splitting: instead of producing a single new particle in the interaction of interest, N particles are produced, each with weight $\frac{1}{N}$. This method may be applied in addition to IB or IF. Storing the whole PG spectrum in stage 1 and sampling it in stage 2 could be viewed as delayed particle splitting. Not implementing the separation between stages 1 and 2 would have been possible by directly generating a PG from the spectra computed in stage 1 as the protons traverse the phantom.

When considering IB, the main disadvantage is the free parameter α, which should be chosen in such a way that sufficiently often a PG is produced. What sufficiently often means will depend on each individual simulation. Moreover, the PG prediction is still a probabilistic process: in some voxels a PG may be produced, in some may not. IF is much more similar to vpgTLE, in that it gives a deterministic prediction per interval. Having a larger number of calls to the physics processes is the main source of overhead for both IB and IF. For example, in an analog Geant4 simulation, we measure that 70% of computing time is spent in nuclear interactions. In a naive implementation of IB or IF, we run through the nuclear interactions an increased number of times, which indicates that the upper limit of the gain is $\frac{1}{70\%} \approx 40\%$. Both IB and IF may benefit from precomputed lookup tables to reduce the time spent performing the additional interactions. Lookup tables are most practical if a single or limited number of outputs are sought, e.g. dose, PG.

Indeed, vpgTLE may be considered as a special optimized case of IF with the following differences:

- the Monte Carlo particle weights are not modified
- whole inelastic processes are precomputed instead of being called at each interaction
- the complete precomputed gamma energy spectrum is stored.

The last two points can also be applied to IF implementations. Hence, PG yield per proton track remains essentially the same between IF and vpgTLE. The difference may essentially be viewed as conceptual, vpgTLE being inspired by the TLE approach originally put forward by Williamson (1987).

As far as the authors of this paper could establish, the only other published variance reduction technique for PGs is described in Sterpin *et al* (2015). This method is fully analytic, incorporating experimental or pretabulated PG emission data and a model that assumes the PG emission region can be modeled similar to the dose in pencil beams. The method ray-traces the materials touched by a pencil beam spot, and computes the expected 1D PG profile by a weighted sum of pregenerated profiles per material, which takes 0.3–10 s. The authors admit that this approach does not deal very well with lateral inhomogeneity, a problem that vpgTLE does not have due to TLE methods in principle being assumption free. Another benefit of vpgTLE is the shape of its output: a 4D image where for each voxel a PG spectrum is recorded. This permits the incorporation of the method in many different kinds of PG detector simulation, not just detectors that measure the range. As a small example, vpgTLE can be used to investigate the origin of the global spectrum as function of depth along the beam path in figure 11. vpgTLE will work with any collimated or Compton camera design, PG spectroscopy, or any future detector that takes advantage of spatial or spectral components. Moreover, vpgTLE does not assume anything about the proton flux: it uses stage 1 to estimate it. The difficulty of estimating PG yields due to beam spots that produce different ranges due to lateral inhomogeneity is therefore avoided: vpgTLE is as sensitive to inhomogeneity as regular MC.

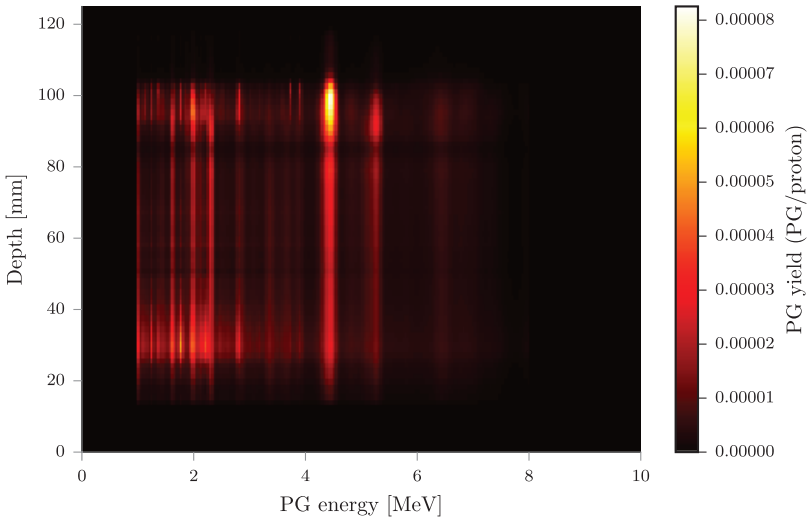

Figure 11. vpgTLE allows one to make full use of spatial and spectral information of PGs. As an illustration, this figure shows a heatmap of PG yields in test case 2 as function of depth and PG energy, integrated over the transverse dimensions. The figure might incline a detector developer to tune a spectral camera to the 4.4 MeV peak.

4.3. Background estimation

vpgTLE does not estimate physical quantities other than PGs produced in the target, which means that no estimate of the background noise can be obtained with this method. For any detector seeking actual application, no analysis is complete without considering the signal-to-background ratio, and methods to improve it. The background consists of tertiary gammas, mainly produced by secondary neutrons undergoing nuclear interactions, in the target but also in elements of the beamline, other objects in the room such as the patient table, and the detector itself (Pinto *et al* 2014). Obtaining the background from simulation is problematic for two reasons:

1. No MC tool correctly estimates the background (see Pinto *et al* (2014) section 3.1.2, Sterpin *et al* (2015) section IV.A.4). Nuclear reaction models are continuously being improved, but as far as we know not specifically for PG *background* at this time.
2. The computing time required for full room simulations is prohibitive, which is why room modeling is left out. Even without room components, the simulation runtime is still long using analog MC, as there are no variance reduction techniques for PG *background*.

Depending on the purpose of the simulation, various ways of dealing with the lack of reliable MC background prediction are considered. Sterpin *et al* (2015), concerned with treatment prediction, deals with the background by adding an offset parameter to their fitting procedure of the PG profiles and accepts the absence of background estimation in their fast PG method. A procedure like this could be applied for other devices. Simulations for camera optimization, effectively signal-to-background optimization for a multitude of configurations, also require a background estimate. Because the background may depend on camera parameters and obtaining a measured background for each configuration is not tractable, most groups limit the number of configurations. The multi-slit camera optimization of Pinto *et al* (2014), which correctly considers hundreds of configurations, was performed with a manual correction of a single background measurement.

Hence, the fact that in current practice the background is measured or modeled separately puts vpgTLE at no particular disadvantage compared to other variance reduction methods or analog MC. However, if physics models could give an accurate estimate of the background, a generalized vpgTLE, most urgently of neutrons and neutron induced PGs, might be a possible avenue for variance reduction of the background estimate. The scorer could be attached to not only the target but also other objects in the treatment room.

4.4. Future improvements

We did not take into account the systematic error (caused by the PGdb) in our analysis. The PGdb is computed by shooting protons of an energy at least as high as used in the treatment plan into a box of a certain material. This means that the PGdb statistics are not as good at low energies as at higher energies, meaning that we have a systematic error around the BP region (where the proton energy is lowest). We might therefore consider supplementing our PGdb with the output of a second, low energy proton beam, simply to reduce the systematic error in the BP range. Since we compute the PGdb once, this has no effect on the efficiency of the vpgTLE method. It might also be possible to fill the database by querying Geant4 for the cross section at the respective bins. We did implement outputs for the analytic systematic and random error output as laid out in El Kanawati *et al* (2015), but stage 2 does not propagate these errors yet, which would be required for a quantitative analysis of the outputs of stage 2. Moreover, we found that this analytic error estimate underestimates with respect to the batch method, because it assumes the independence of the track lengths and proton energies, a type of problem IF or IB techniques would not have. Therefore, at this time, the variance of vpgTLE can only be obtained by employing the batch method.

A thorough analysis of the sparseness of the PGyd has not been conducted, but there is a likely opportunity for memory optimization here. Another option is to reduce the dimensionality of the image to that which the user requires. Users investigating their collimated camera may for example set the PG spectrum to a narrower window and a coarser binning, reducing memory consumption accordingly.

Precomputing an effective linear production coefficient, the main principle of vpgTLE, could be performed for other particles. Adding for instance an effective linear neutron production coefficient may supplement the vpgTLE output with a correct estimate of neutron-induced gamma noise in a PG detector, giving an indication of the background, not just the PG signal. However, such an addition must be investigated to obtain the real efficiency.

5. Conclusion

vpgTLE is a generic drop-in alternative for computing the expected PG output in voxelized geometries. The method has a fixed memory requirement (a 4D image) with a typical memory size of the order of a few hundred megabytes. The method reaches a global gain factor of 10^3 for a clinical CT image and treatment plan with respect to analog MC. A median convergence of 2% for the most distal energy layer is reached in approximately 4 h on a single core, at which point the output has stabilized to within 10^{-4} of an analog reference simulation, when the range or the spectrum is considered. The authors think the method will be of interest to those developing and simulating PG detection devices, as well as clinicians studying complex clinical cases that require the precision and accuracy of MC level simulations not offered by analytic algorithms.

The vpgTLE method is open source and fully integrated in Gate. It is available from release 7.2 onwards.

Acknowledgments

This work was partly supported by Labex PRIMES ANR-11-LABX-0063, t-Gate ANR-14-CE23-0008, France Hadron ANR-11-INBS-0007 and LYric INCa-DGOS-4664. We thank Denis Dauvergne for his assistance with preparing the copy for this paper and Marc Verderi of the Geant4 collaboration for discussions on variance reduction in Geant4.

References

Biegun A K *et al* 2012 Time-of-flight neutron rejection to improve prompt gamma imaging for proton range verification: a simulation study *Phys. Med. Biol.* **57** 6429–44

El Kanawati W, Létang J M, Dauvergne D, Pinto M, Sarrut D, Testa É and Freud N 2015 Monte Carlo simulation of prompt γ-ray emission in proton therapy using a specific track length estimator *Phys. Med. Biol.* **60** 8067–86

Golnik C *et al* 2014 Range assessment in particle therapy based on prompt γ-ray timing measurements *Phys. Med. Biol.* **59** 5399–422

Gueth P, Dauvergne D, Freud N, Létang J M, Ray C, Testa E and Sarrut D 2013 Machine learning-based patient specific prompt-gamma dose monitoring in proton therapy *Phys. Med. Biol.* **58** 4563–77

Henyey L G and Greenstein J L 1941 Diffuse radiation in the galaxy *Astrophys. J.* **93** 70–83

Jan S *et al* 2004 GATE: a simulation toolkit for PET and SPECT *Phys. Med. Biol.* **49** 4543

Janssen F M F C, Landry G, Cambraia Lopes P, Dedes G, Smeets J, Schaart D R, Parodi K and Verhaegen F 2014 Factors influencing the accuracy of beam range estimation in proton therapy using prompt gamma emission *Phys. Med. Biol.* **59** 4427–41

Knopf A C and Lomax A 2013 *In vivo* proton range verification: a review *Phys. Med. Biol.* **58** R131–60

Kurosawa S *et al* 2012 Prompt gamma detection for range verification in proton therapy *Curr. Appl. Phys.* **12** 364–8

Lopes P C *et al* 2015 Time-resolved imaging of prompt-gamma rays for proton range verification using a knife-edge slit camera based on digital photon counters *Phys. Med. Biol.* **60** 6063

Marcatili S, Villoing D, Garcia M P and Bardiès M 2014 Multi-scale hybrid models for radiopharmaceutical dosimetry with Geant4 *Phys. Med. Biol.* **59** 7625–41

Min C H, Kim C H, Youn M Y and Kim J W 2006 Prompt gamma measurements for locating the dose falloff region in the proton therapy *Appl. Phys. Lett.* **89** 38–41

Moteabbed M, España S and Paganetti H 2011 Monte Carlo patient study on the comparison of prompt gamma and PET imaging for range verification in proton therapy *Phys. Med. Biol.* **56** 1063–82

Parodi K, Pönisch F and Enghardt W 2005 Experimental study on the feasibility of in-beam PET for accurate monitoring of proton therapy *IEEE Trans. Nucl. Sci.* **52** 778–86

Perali I *et al* 2014 Prompt gamma imaging of proton pencil beams at clinical dose rate *Phys. Med. Biol.* **59** 5849–71

Pinto M *et al* 2015 Absolute prompt-gamma yield measurements for ion beam therapy monitoring *Phys. Med. Biol.* **60** 565–94

Pinto M, Dauvergne D, Freud N, Krimmer J, Létang J M and Testa E 2016 Assessment of Geant4 prompt-gamma emission yields in the context of proton therapy monitoring *Front. Oncol.* **6** 1–7

Pinto M, Dauvergne D, Freud N, Krimmer J, Letang J M, Ray C, Roellinghoff F and Testa E 2014 Design optimisation of a TOF-based collimated camera prototype for online hadrontherapy monitoring *Phys. Med. Biol.* **59** 7653–74

Richter C *et al* 2016 First clinical application of a prompt gamma based *in vivo* proton range verification system *Radiother. Oncol.* **118** 232–7

Robert C *et al* 2013 Distributions of secondary particles in proton and carbon-ion therapy: a comparison between GATE/Geant4 and FLUKA Monte Carlo codes *Phys. Med. Biol.* **58** 2879–99

Roellinghoff F *et al* 2011 Design of a Compton camera for 3D prompt-γ imaging during ion beam therapy *Nucl. Instrum. Methods Phys. Res.* A **648** S20–3

Roellinghoff F *et al* 2014 Real-time proton beam range monitoring by means of prompt-gamma detection with a collimated camera *Phys. Med. Biol.* **59** 1327–38

Sarrut D *et al* 2014 A review of the use and potential of the GATE Monte Carlo simulation code for radiation therapy and dosimetry applications *Med. Phys.* **41** 064301

Sheldon E and Van Patter D M 1966 Compound inelastic nucleon and gamma-ray angular distributions for even- and odd-mass nuclei *Rev. Mod. Phys.* **38** 143–86

Smeets J *et al* 2012 Prompt gamma imaging with a slit camera for real-time range control in proton therapy *Phys. Med. Biol.* **57** 3371–405

Sterpin E, Janssens G, Smeets J, Stappen F V, Prieels D, Priegnitz M, Perali I and Vynckier S 2015 Analytical computation of prompt gamma ray emission and detection for proton range verification *Phys. Med. Biol.* **60** 4915–46

Testa E, Bajard M, Chevallier M, Dauvergne D, Foulher F L, Poizat J C, Ray C, Testa M, Freud N and Létang J M 2008 Monitoring the Bragg peak location of 73 MeV/u carbon ions by means of prompt γ-ray measurements *Appl. Phys. Lett.* **93** 093506

Testa M, Verburg J M, Rose M, Min C H, Tang S, Bentefour E H, Paganetti H and Lu H M 2013 Proton radiography and proton computed tomography based on time-resolved dose measurements *Phys. Med. Biol.* **58** 8215–33

Verburg J M and Seco J 2014 Proton range verification through prompt gamma-ray spectroscopy *Phys. Med. Biol.* **59** 7089–106

Verburg J M, Riley K, Bortfeld T and Seco J 2013 Energy- and time-resolved detection of prompt gamma-rays for proton range verification *Phys. Med. Biol.* **58** L37–49

Verburg J M, Shih H A and Seco J 2012 Simulation of prompt gamma-ray emission during proton radiotherapy *Phys. Med. Biol.* **57** 5459–72

Williamson J F 1987 Monte Carlo evaluation of kerma at a point for photon transport problems *Med. Phys.* **14** 567–76